Chapter 9

Lexical semantic change for Ancient Greek and Latin

Valerio Perrone[a], Simon Hengchen[b], Marco Palma[c], Alessandro Vatri[d], Jim Q. Smith[c,e] & Barbara McGillivray[f,e]

[a]Amazon [b]University of Gothenburg [c]University of Warwick [d]University of Oxford [e]The Alan Turing Institute [f]University of Cambridge

Change and its precondition, variation, are inherent in languages. Over time, new words enter the lexicon, others become obsolete, and existing words acquire new senses. Associating a word with its correct meaning in its historical context is a central challenge in diachronic research. Historical corpora of classical languages, such as Ancient Greek and Latin, typically come with rich metadata, and existing models are limited by their inability to exploit contextual information beyond the document timestamp. While embedding-based methods feature among the current state of the art systems, they are lacking in their interpretative power. In contrast, Bayesian models provide explicit and interpretable representations of semantic change phenomena. In this chapter we build on GASC, a recent computational approach to semantic change based on a dynamic Bayesian mixture model. In this model, the evolution of word senses over time is based not only on distributional information of lexical nature, but also on text genres. We provide a systematic comparison of dynamic Bayesian mixture models for semantic change with state-of-the-art embedding-based models. On top of providing a full description of meaning change over time, we show that Bayesian mixture models are highly competitive approaches to detect binary semantic change in both Ancient Greek and Latin.

1 Introduction

The study of lexical semantics in a diachronic perspective is of primary importance in lexicography, historical linguistics and other humanities. Capturing the

semantic spectrum and historical change of individual words as well as performing large-scale diachronic analyses of the lexicon can help us answer important questions about the development of our culture and heritage. Recent research in natural language processing (NLP) has led to the development of computational models of lexical semantic change (LSC) which have the potential to add new insights to diachronic semantics. Most computational research in this area, however, has focussed on extant languages, and only a few attempts have been made to tackle this topic for ancient languages.

To address this, Perrone et al. (2019) introduced GASC (genre-aware semantic change), a novel dynamic Bayesian mixture model for semantic change, where the evolution of word senses over time is based on distributional information and on additional features, specifically genre. GASC can decouple sense probabilities and genre prevalence, a critical task in the case of genre-unbalanced languages corpora, and can incorporate different categorical metadata, such as author, geography, or style. GASC was developed for Ancient Greek and represents the state-of-the-art in computational modelling of lexical semantic change for this language.

On the other hand, word-embedding models have become the most common methods adopted in lexical semantic change detection (Kutuzov et al. 2018) and an open question remains regarding which methods are most appropriate for ancient languages. In this chapter, we offer the first systematic evaluation of Bayesian dynamic mixture models and word embeddings models for semantic change in Latin and Ancient Greek. These ancient languages provide insightful test cases of automatic lexical semantic change for several reasons. First, as in many other languages, a large number of Latin and Ancient Greek words are polysemous (Clarke 2010), and polysemous words offer us a chance to study semantic variation, particularly across genres, and its relation to semantic change (Leiwo et al. 2012). Also, the literary traditions of these two languages have rich transcribed high-quality corpora covering a large number of literary genres. Moreover, they offer the opportunity to test the performance of different methods on use data spanning several centuries. Finally, we can rely on the scholarship of these languages to validate our computational systems. Our code is freely available at https://git.io/Jqe7U.

The word *mus* is an example of polysemous word (it can mean 'mouse', 'mussel' or 'muscle'). The variation in the distribution of meanings over time per genre is displayed in Figure 9.1. In this graph, lines represent the percentage of the occurrences of the target word in a literary genre across centuries, while bars represent the percentage of the occurrences of a specific sense of *mus* across centuries. When the trend in any line agrees with the one for any set of bars (for
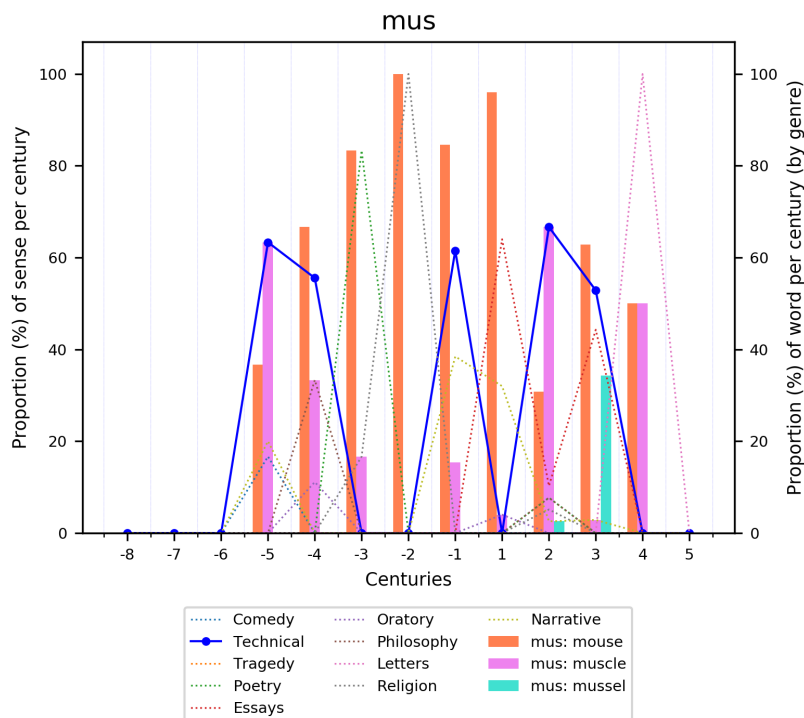
Figure 9.1: Distribution of *mus* 'mouse'/'muscle'/'mussel' by genre vs
its senses over time (Perrone et al. 2019). Lines track *mus* proportions
in each genre and century, while bars show the *mus* occurrence pro-
portions with each sense and century.

instance, the distribution of 'muscle' over time tracks the blue line correspond-
ing to the distribution of *mus* in technical genres), there might be evidence of
genre-related changes.

In technical texts, we expect polysemous words to have a technical sense
('muscle' in the case of *mus*). On the other hand, in works more closely repre-
senting general language (comedy, oratory, historiography) we expect words to
appear in their more concrete and less metaphorical senses ('mouse' or 'mussel'
in the case of *mus*), although we cannot always assume that the same distribution
holds in a number of other genres, such as philosophy and tragedy.

2 Related work

In recent years, NLP research has made great advances in the area of semantic
change detection and modelling, with methods ranging from topic-based mod-

els (Boyd-Graber et al. 2007, Cook et al. 2014, Lau et al. 2014, Wijaya & Yen-iterzi 2011, Frermann & Lapata 2016), to graph-based models (Mitra et al. 2014, 2015, Tahmasebi & Risse 2017), and word embeddings (Kim et al. 2014, Basile & McGillivray 2018, Kulkarni et al. 2015, Hamilton et al. 2016, Dubossarsky et al. 2017, Tahmasebi 2018, Rudolph & Blei 2018, Jatowt et al. 2018, Dubossarsky et al. 2019), to cite but a few.[1] However, models used in previous work are purely based on words' lexical distribution information and do not account for language variation features such as text type or genre. One reason for this is that genre-balanced corpora (such as COHA in Davies 2012) or single-genre corpora (such as newspapers, or Twitter, cf. e.g. Shoemark et al. 2019) are typically used. However, the strong role played by such factors in determining the sense of a word in context has been acknowledged in NLP research at least since Gale et al. (1992)'s idea of "one sense per discourse", according to which polysemous words tend to display the same sense in the same discourse. This principle has been widely adopted in word sense disambiguation research, with some more recent adaptations such as "one sense per Wikipedia category" (Scarlini et al. 2020).

Semantic change in ancient languages, especially on a large scale and over a long time period, is an under-explored research area. Previous work has mainly been qualitative in nature, due to the complexity of the phenomenon (cf. e.g. Leiwo et al. 2012, Clackson 2011). Some work has been done on training word embeddings on Ancient Greek (Rodda et al. 2019) and Latin (Sprugnoli et al. 2019) corpora, but not in a diachronic perspective. With the exception of a few works (Bamman & Crane 2011, Eger & Mehler 2016, Rodda et al. 2016, Perrone et al. 2019, McGillivray et al. 2019), two of which this chapter is based on and completes, no previous work has focussed on ancient languages.[2]

Recent work on languages other than English is rare but exists: Falk et al. (2014) use topic models to detect changes in French and Hengchen (2017) uses similar methods to tackle Dutch. Cavallin (2012) and Tahmasebi (2018) focus on Swedish, with the comparison of verb-object pairs and word embeddings, respectively. Zampieri et al. (2016) use SVMs to assign a time period to text snippets in Portuguese, and Tang et al. (2016) work on Chinese newspapers using S-shaped models. Most work in this area focusses on simply detecting the occurrence of semantic change, while Frermann & Lapata (2016)'s system, SCAN, takes into

---

[1]For an overview of the NLP literature, we refer to Tahmasebi et al. (2018), and Kutuzov et al. (2018) for a focus on neural embeddings. For an overview of the existing challenges in modelling and detecting semantic change, we refer to Hengchen et al. 2021 [this volume].

[2]To this list, we add the very recent SemEval 2020 Task 1 shared task on unsupervised lexical semantic change detection (Schlechtweg et al. 2020), which had Latin as one of its four target languages.

account synchronic polysemy and models how the different word senses evolve across time. More recently French has been further tackled by Jawahar & Seddah (2019), Frossard et al. (2020) and Montariol & Allauzen (2020), and German has been the focus of extensive work (Schlechtweg et al. 2017, 2018, 2019, 2020).

Our work bears important connections with the topic model literature. The idea of enriching topic models with document-specific author meta-data was explored in Rosen-Zvi et al. (2004) for the static case. Several time-dependent extensions of Bayesian topic models have been developed, with a number of parametric and nonparametric approaches (Blei & Lafferty 2006, Rao & Teh 2009, Ahmed & Xing 2012, Dubey et al. 2013, Perrone et al. 2017). In this chapter, we transfer such ideas to semantic change, where each datapoint is a bag of words associated to a single sense (rather than a mixture of topics). Excluding cases of intentional ambiguity, which we expect to be rare, we assume that there are generally no ambiguities in a context, and each word instance maps to a single sense. We acknowledge that this assumption can be seen as going against historical semantics literature (e.g. Traugott & Dasher 2001) which states that variation in context is the seed of semantic change.

3 The corpora

In order to conduct our experiments, we made use of two large diachronic corpora of Latin and Ancient Greek: LatinISE (McGillivray & Kilgarriff 2013) for Latin and the Diorisis Annotated Ancient Greek Corpus (Vatri & McGillivray 2018) for Ancient Greek. Our models require genre information. Genre-annotated corpora are not particularly common in NLP, where most tasks rely on specific genres (e.g. Twitter) or on genre-balanced corpora such as COHA (Davies 2002), but they are more prevalent within the humanities, and especially Classics. Additionally, research on automated genre identification has been flourishing for decades (e.g. Kessler et al. 1997), making the need for genre information in a potential corpus not as much of a hindrance as it could be thought.[3]

The Diorisis Annotated Ancient Greek Corpus contains 820 texts spanning between the beginnings of the Ancient Greek literary tradition (8th century BCE) and the 5th century CE. It is lemmatized and part-of-speech-tagged and contains 10,206,421 word tokens. Diorisis is the largest openly available annotated corpus

---

[3]While the influence of genre has been extensively studied in historical linguistics (see, for example, the extensive work by Biber & Finegan 1989), we use in this chapter a slightly different notion of *genre*: literary genre, as defined by classicists.

of Ancient Greek. The corpus covers a number of Ancient Greek literary and technical genres: poetry (narrative, choral, epigrams, didactic), drama (tragedy, comedy), oratory, philosophy, essays, narrative (historiography, biography, mythography, novels), geography, religious texts (hymns, Jewish and Christian Scriptures, theology, homilies), technical literature (medicine, mathematics, natural science, tactics, astronomy, horsemanship, hunting, politics, art history, rhetoric, literary criticism, grammar), and letters.

The LatinISE corpus (McGillivray & Kilgarriff 2013) covers 1,274 texts from between the beginnings of the Latin literary tradition (2nd century BCE) and the contemporary era (21st century CE). It has been automatically lemmatized and part-of-speech tagged. A domain expert manually added genre information for the following genres: comedy, essays, law, letters, narrative, oratory, philosophy, poetry, Christian, technical, tragedy. All Christian writings (including letters and poems) were assigned the genre Christian. This excludes philosophical but not theological or ecclesiological treatises composed by Christian writers.

4 Bayesian semantic change models

4.1 Domain knowledge elicitation

While NLP provides powerful tools to analyse texts, a central challenge is to ensure that outputs are explainable and that new discoveries can be placed within the context of the current state of the art in specific disciplines where NLP methods are applied. Bayesian methods have proved very useful within scientific modelling to incorporate domain explanations. In the Bayesian setting, expert judgements can be embedded directly into a probabilistic framework in the form of a prior. For instance, if historians know that a certain sense was popular in a given century, this information can be directly encoded into the model by changing the prior probability distribution for that sense. Data can then be analysed from these belief statements and a prior to posterior analysis performed, which helps domain experts adjust their beliefs in the light of the new available information (see for example Smith 2010, O'Hagan & Oakley 2014). These new outputs will be consistent with the explanations embedded within the probabilistic model, making results interpretable.

The challenge of applying Bayesian reasoning within the humanities is that typically domain experts have not been trained to reason probabilistically. Therefore, it is not possible to ask domain experts to provide direct probabilistic inputs to the Bayesian model. What it is possible instead is to elicit structural information, which can take a wide range of forms depending on the domain (Wilkerson

& Smith 2018). These structural models can usually be represented by a graph (i.e., as a set of nodes and connecting arcs) which captures the fundamental entities and their relationships. For example, an expert may know that a certain author predominantly uses *mus* to mean 'mouse'. The Bayesian modeller can then simply introduce a new node representing the author and condition the probability of using senses to the author variable. Once the graphical model is in place, we let the data quantify a joint probability model.

This work leverages the Bayesian network, one of the most developed structural models of this type. This structure embeds simple assertions about what measurements might be informative, in a way described by Korb & Nicholson (2009), Smith (2010), and Pearl (2009). Working backwards from the properties of the object of interest, we produce sequentially a collection of direct and indirect influences across the whole domain. The composite of the relationships can then be expressed by a single graph, called a plate diagram (see Figure 9.2 for such a plate diagram of our model). This plate diagram determines the factorisation of the corresponding probability density over these measurements.

We aimed to apply these structural elicitation techniques to the study of semantic variation and change in Latin and Ancient Greek. From discussions with Ancient Greek and Latin experts who have extensive experience with the corpora at hand, it emerged that one of the main drivers of this variation was the particular genre of the text. For instance, in works more closely related to general language (i.e. non-specialised, or non purely poetic language), such as comedy or historiography, we expect words to appear in their concrete and less metaphorical senses. The Ancient Greek word *mus* within a technical text would more likely mean 'muscle', while in narrative texts the meaning would more likely be 'mouse'. Such variations were believed to abound within the studied corpus. Since both the genre of texts was known to vary over time and text preservation to the current date depended on genre, any analysis which ignored genre might deduce a spurious change in overall meaning simply explainable from drifts in genre and selection effects influencing preservation. Having elicited this domain judgement, it was clear how to proceed. We simply modified the structure of our Bayesian model by adding genre as an additional observable variable (or node in the plate diagram). Conditioning on the observed genre, we could then have a specific distribution over senses accounting for genre-specific word usage patterns. Details of the model are given in the next section.
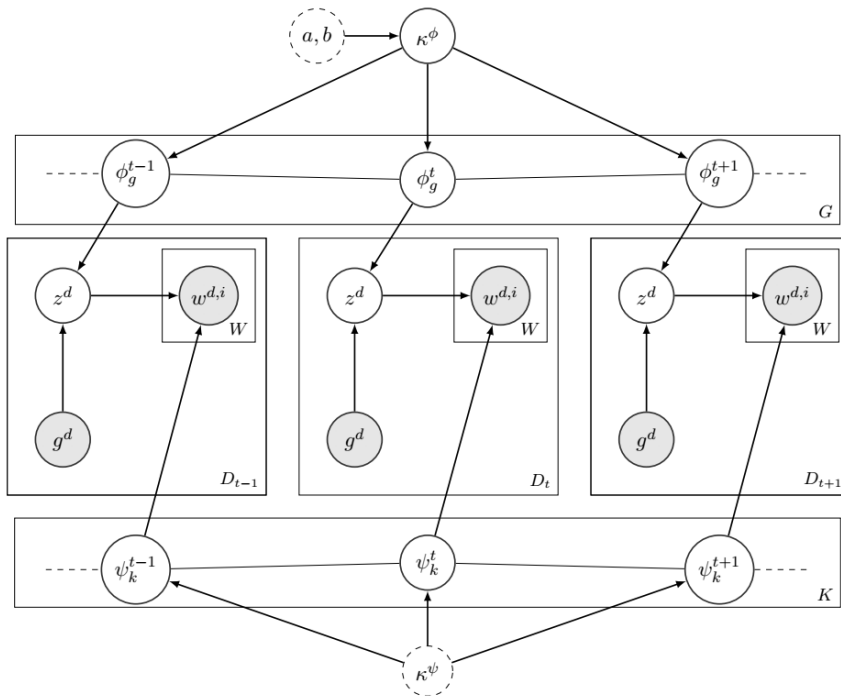
Figure 9.2: GASC plate diagram with three time periods (Perrone et al. 2019).

4.2 Genre-aware semantic change

A successful approach to model semantic change in Ancient Greek is GASC (Perrone et al. 2019). The starting point is a lemmatised corpus pre-processed into a set of text snippets of size W, each containing an instance of the word under study (referred to as "target word" in the remainder). The inferential task is to detect the sense associated to the target word in the given context, and describe the evolution of sense proportions over time.

We briefly summarise the generative model for GASC (illustrated by the plate diagram in Figure 9.2), which extends SCAN (Frermann & Lapata 2016) to be genre-aware and is described in detail in Perrone et al. (2019). First, suppose that throughout the corpus the target word is used with K different senses, where we define a sense at time t as a distribution ψ_k^t over words from the dictionary. This statistical definition of sense is necessary to formalize the generative models presented in this work, and will be used throughout the rest of the paper.[4] These

---

[4]We follow the terminology adopted by Frermann (2017) and represent the meaning of a word

distributions are used to generate text snippets by drawing each of their words from the dictionary based on a multinomial distribution. Based on the intuition that each genre is more or less likely to feature a given sense, we assume that each of G possible text genres determines a different distribution over senses. Each observed document snippet is then associated with a genre-specific distribution over senses $\phi_{g^d}^t$ at time t, where g^d is the observed genre for document d. Conditioning on the observed genre yields a specific distribution over senses accounting for genre-specific word usage patterns. Word and sense distributions evolve over time with Gaussian changes, allowing for smooth transitions.

The model can be applied to different inferential goals: we can focus on the evolution of sense probabilities or on the changes within each sense. As we define a sense at time t as a probability distribution over words from the dictionary, this means that we can either choose to focus on the change of the sense probability over time or on the change in probability of the words characterising that sense. For each of these aims, we can use several hyperparameter combinations for K^ϕ, which is drawn from the prior distribution as determined by a and b, and K^ψ. To effectively detect semantic change points, the sense probabilities should not vary too smoothly over time and the bag of words should remain stable throughout the time periods.[5] For these reasons, we set the hyperparameters $a = b = 1$, $K^\psi = 100$ (equivalent to Setting 3 in Perrone et al. 2019). In particular, the hyperparameter K^ψ controls the homogeneity of the bag of words within the same sense and allows the emergence of new senses. This hyperparameter setting is used for SCAN and GASC on Latin, as well as for SCAN on Ancient Greek. For GASC on Ancient Greek, where the corpus size and the number of occurrences of target words is split between genres, the set of hyperparameters used is $a = 7, b = 3$, $K^\psi = 10$, as in Frermann & Lapata (2016).

Further quantities to be set before running the Bayesian models are the number of iterations and the window size parameter W. The first runs of the Bayesian models usually show high variability in the results before convergence occurs; therefore, it is necessary to use a large number of iterations, especially for small sample sizes. For posterior inference we discard the first 100 iterations (burn-in period) and we run 2,500 iterations for models on Latin and 10,000 for models on

---

as a set of senses, each of which captures "an internally coherent aspect of its meaning, and is characterized through a set of words that are associated with that sense" (Frermann 2017: 173). We also assume that each instance of a target word in the corpus refers to one and only one sense.

[5]We acknowledge that the task of detecting change points is a drastically reduced view of semantic change. Nonetheless, as further explained in Section 6, this is required for ground truth evaluation.

Ancient Greek. The window size parameter W, namely the number of words to the left and right of an instance of the target, must also be carefully chosen not to introduce noisy irrelevant contextual words. Following Frermann & Lapata (2016) and Perrone et al. (2019), we fix the window size W to 5 for all methods and languages.

For posterior inference, we extend the blocked Gibbs sampler proposed in Frermann & Lapata (2016). The full conditional is available for the snippet-sense assignment, while to sample the sense and word distributions we adopt the auxiliary variable approach from Mimno et al. (2008). The sense precision parameters are drawn from their conjugate Gamma priors. For the distribution over genres we proceed as follows. First, sample the distribution over senses ϕ_g^t for each genre $g = 1, \dots, G$ following Mimno et al. (2008). Then, sample the sense assignment conditioned on the observed genre from its full conditional:

$$p(z^d \mid g^d, \mathbf{w}, t, \phi, \psi) \propto p(z^d \mid g^d, t) p(\mathbf{w} \mid t, z^d) = \phi_g^t \prod_{w \in \mathbf{w}} \psi_w^{t, z^d}.$$

This setting easily extends to sample genre assignments for tasks where, for example, some genre metadata are missing.

5 Embedding-based models

Neural-based word vectors are currently the most used representations in LSC. While skip-gram with negative sampling (SGNS, Mikolov et al. 2013) type embeddings have the limitation that they conflate senses of a word to a single vector representation, they currently perform better than other approaches, including contextual models such as ELMO (Peters et al. 2018) and BERT (Devlin et al. 2019), as reported by Schlechtweg et al. (2020).

In this chapter, we compare GASC and SCAN to the current state of the art in LSC (temporal referencing (TR), Dubossarsky et al. 2019), as well as with the oft-used combination of independently-trained SGNS models that are subsequently aligned using Orthogonal Procrustes (OP) proposed by Hamilton et al. (2016). Both models are very similar and rely on the same algorithm with the difference that TR, in which target words have different representations for every time bin but context words do not, has repeatedly been shown to produce much less noisy models (e.g. in Cassotti et al. 2020, Zamora-Reina & Bravo-Marquez 2020). In order to compare their performance with GASC, we train models on the whole corpus ("NAIVE"), as well as on genre subcorpora. For Ancient Greek we train models on Technical, NOT-technical, Narrative, NOT-narrative subcorpora, while Latin is divided between Christian and NOT-Christian.

6 Evaluation

Evaluating models tackling lexical semantic change is notoriously challenging. Schlechtweg et al. (2020) present the first shared task on unsupervised lexical semantic change detection, organized as part of the SemEval 2020 workshop. The task focusses on two subtasks: a binary classification task (for a set of target words, decide which words lost or gained senses between a time period t_1 and a time period t_2) and a ranking subtask (rank the same set of target words according to their degree of lexical semantic change between t_1 and t_2). The task provides gold standard sets for three extant languages (English, German, and Swedish) and one extinct language (Latin).

The Latin gold standard reflects the lexical semantic change in a portion of the Latin lexicon from Before the Common Era (BCE) and the Common Era (CE). For each of 40 lemmas selected from the corpus, expert annotators annotated 30 sentences extracted from a subcorpus of LatinISE consisting of texts from BCE, and 30 sentences from CE. For each sentence, the annotators selected one of four values (4 – identical, 3 – closely related, 2 – distantly related, 1 – unrelated) for each dictionary sense of the lemma, indicating the degree of similarity between the usage of the lemma reflected in the sentence and the dictionary sense. This choice of design implying that every target word has a closed set of possible senses corresponding to those listed in their respective dictionary entries is justified in the original paper.

The annotated data was analysed with a clustering technique that identified 26 lemmas as "changed" lemmas (meaning that they underwent lexical semantic change between BCE and CE) and 14 lemmas as "unchanged" (meaning that they did not undergo lexical semantic change). For details on the clustering and the annotation, see Schlechtweg et al. (2020) and Schlechtweg et al. (2021). The SemEval task competition and the subsequent article describing a subset of the systems that took part in it offers the first systematic evaluation of state-of-the-art systems for automatic lexical semantic change detection.

Word embedding models build vector representations of a word for every time slice at hand. For two time intervals t_1 and t_2, we then use a similarity measure (usually, cosine similarity) as a proxy to determine the semantic change between the vectors w_{t_1} and w_{t_2} for a specific word between these time slices:

$$\text{cosine\_similarity}(w_{t_1}, w_{t_2}) = \frac{w_{t_1} \cdot w_{t_2}}{\|w_{t_1}\|\|w_{t_2}\|},$$

where $\|\cdot\|$ denotes the Euclidian norm. A high cosine similarity (e.g., close to 1) means no difference for word w between time slices t_1 and t_2, and a low cosine similarity indicates a high difference.

As our ground truth consists of a binary classification (no-change/change, cf. Section 7), we must transform the cosine similarity value, bounded between -1 and 1, into a decision. While manual thresholding on the cosine is usually applied, recent work (Zhou & Li 2020) shows that determining the threshold in a data-driven way is beneficial. We thus follow prior work on Latin and fit a Gamma distribution of the cosine similarities for all target words between t_0 and t_1, and consider every cosine similarity below the 75-quantile value as the threshold for a change decision.[6]

On the other hand, dynamic Bayesian mixture models, such as SCAN and GASC, are designed to infer the smooth evolution of sense probabilities over time. We adapt these methods to detect sense change points as follows. First, we compute the mean and standard deviations of the posterior sense probabilities over time based on the Gibbs samples obtained during inference. Then, we infer that there has been a significant drop or rise of a sense if its posterior mean probability changes by at least two standard deviations over time. In case of a significant drop we infer that a sense disappeared, and in case of a significant increase we infer that a new sense appeared in the data. If sense probabilities do not change significantly over time, we conclude that no meaning change occurred. Note that, unlike SCAN, GASC outputs a sense probability over time for each genre, and we thus check across all genres whether a significant change of sense probability occurred over time. While we adopt this approach for simplicity, change point analysis has been studied extensively in the context of Gaussian dynamic state space models. We refer to West & Harrison (1997) and Frühwirth-Schnatter (2006) for more sophisticated approaches to detect change points, which also allow for returning a probability distribution over change points.

7 Experiments

7.1 Tasks and baselines

We compared SCAN and GASC to a wide range of baselines on the task of detecting binary change in both Latin and Ancient Greek. Perrone et al. (2019) and Vatri et al. (2019) present a gold standard set created for the purpose of evaluating GASC on Ancient Greek. This set consists of the sense annotation of corpus sentences for three words (*mus* 'mouse'/'muscle'/'mussel', *harmonia* 'fastening'/ 'agreement'/'musical scale, melody', *kosmos* 'order'/'world'/'decoration'). These lemmas display a high degree of clear-cut polysemy,[7] especially across genres

---

[6]We thank Jinan Zhou and Jiaxin Li for providing us with their implementation.

[7]By clear-cut polysemy, we mean that the different senses of a word are not strongly related.

(Liddell et al. 1996, Pollitt 1974), and were chosen as "non-changed" words. We considered two additional lemmas, which display a degree of lexical semantic change in the time period under study, *parabole* 'comparison'/'parable' and *paradeisos* 'garden'/'paradise' (McGillivray et al. 2019). *Paradeisos* is an Avestan loan word that first appeared in Greek in the fifth century BCE to indicate a 'royal park' and probably became common after the Macedonian conquest of the Persian empire. This word was chosen by the Greek translators of the Pentateuch to refer to the garden of Eden around the third century BCE (Kyrtatas 2007). The meaning of *parabole*, in turn, specialized from that of 'comparison' to that of 'short moral narrative' with the New Testament (first century CE). For Latin, we made use of the SemEval task's gold standard, consisting of 26 "changed" lemmas and 14 "non-changed" lemmas between BCE and CE. We start by visualizing the smooth semantic change inferred by GASC, and then compare the ability of dynamic Bayesian mixture models to detect binary semantic change with the state of the art, both on Latin and Ancient Greek.
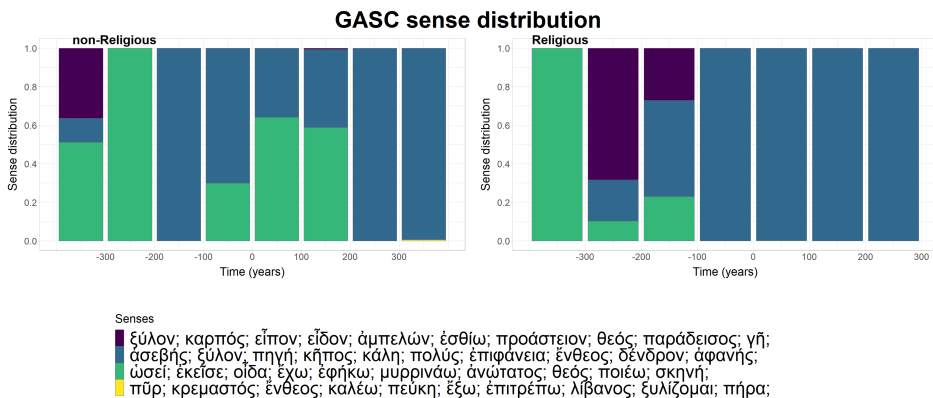
7.2 Smooth semantic change

Figure 9.3: Semantic change in Ancient Greek. Visualization of the probability distributions produced by GASC on the Religious genre for the word *paradeisos* ('garden'/'paradise'). Negative numbers refer to years BCE.

Dynamic Bayesian mixture models are able to infer the full evolution of sense probabilities over time. In particular, GASC is able to do so for each genre provided as input. Figure 9.3 shows the time distribution of the senses of *paradeisos* outputted by GASC run on the Religious vs. non-Religious genres. The

four senses identified by GASC may be interpreted as identifying the meaning 'garden' (senses 3-green and 4-yellow), and 'garden of Eden/(Biblical) paradise' (senses 1-purple and 2-blue). The two senses are not easily distinguishable (since the Biblical paradise is described as a physical garden) and all senses share a number of words, including, notably, *theos* 'God' or the derived adjective *entheos* 'inspired by God'. However, the first two senses contain a number of words that are easily identifiable as connected to the Biblical narration of the fall of man (e.g. *karpos*, (the forbidden) 'fruit' and *esthio*, 'eat') while the remaining senses suggest references to other proverbial gardens (e.g. *kremastos* 'hanging' garden of Babylon). The diachronic evolution of sense distributions in the plots shows that the Biblical meaning comes to rise around the third century BCE in religious texts, which corresponds precisely to the beginning of the translation of the Bible in Greek, and will prevail throughout the Christian era. The graph displaying the computed distribution of senses in non-religious genres captures well the fact that between the first century BCE and the second century CE *paradeisos* is attested a number of times in the works of historians and geographers represented in the corpus. After the third century, this word is very rarely attested in the works included in the Diorisis corpus and almost half of its occurrences in non-religious texts refer to the Biblical garden of Eden.

7.3 Binary semantic change

Next, we evaluated the ability to recover ground truth about binary semantic change on both Latin and Ancient Greek. For Latin, we recall that ground truth consists of 40 target lemmas, 26 of which underwent semantic change. We ran the genre-aware baselines by specifying whether a text belongs to the Christian genre or not. Results in Table 9.1 show that Bayesian models are highly competitive with the best baseline obtained in the SemEval task, with SCAN achieving the highest F1 score. This is striking as dynamic Bayesian mixture models are designed for capturing smooth semantic change over time, rather than binary semantic change across a pair of time points. In addition, only focusing on non-Christian genres decreases the recall of SGNS and TR. This is expected as the 26 lemmas that underwent semantic change did so due to the rise of a new Christian meaning.

We then evaluated each method on Ancient Greek, further adapting SGNS and TR to use genre information and focus on technical and narrative texts. To evaluate GASC, we use Religious as the genre for *parabole* and *paradeisos*, while Technical and Narrative for *mus*, *harmonia* and *kosmos*, with results being averaged across the five words. Results are shown in Table 9.3. While the small number of

target words makes these results mainly illustrative, dynamic Bayesian mixture models emerge as competitive approaches. Consistently with Latin, GASC and SCAN outperform most baselines. To better understand how differently SCAN and TR (the two best-perfoming systems) behave, we refer to the confusion matrix in Table 9.2.

Table 9.1: Semantic change in Latin. Comparison of SCAN and GASC with SGNS, TR and the best baseline from the SemEval task. Results in terms of precision, recall, and F1-score ("F1") averaged across all 40 available words. Results for TR_NAIVE are by Zhou & Li (2020).

Latin (BCE/CE)	Precision	Recall	F1 score
SCAN	0.684	1.000	**0.813**
GASC	0.650	0.920	0.762
SGNS_NOT-christian	1.000	0.308	0.471
SGNS_NAIVE	0.900	0.347	0.500
TR_NOT-christian	0.667	0.231	0.343
TR_NAIVE	0.769	0.769	0.769
Best baseline	0.650	1.000	0.788

Table 9.2: Confusion matrix for binary change in Latin for SCAN and Temporal Referencing. TP = true positive, TN = true negative, FP = false positive, FN = false negative.

System	TP	TN	FP	FN
SCAN	26	2	12	0
TR_NAIVE	20	8	6	6

8 Discussion and conclusion

This work investigates semantic change in Latin and Ancient Greek through several state-of-the-art models. We adapted, discussed and applied a number of algorithms to the case of ancient languages. The adoption of quantitative corpus-based approaches in historical linguistics is growing (Jenset & McGillivray 2017). However, computational approaches to lexical semantic change detection have not yet been widely used in historical linguistics research (McGillivray 2020),

Table 9.3: Semantic change in Ancient Greek. Comparison of SGNS, TR, GASC and SCAN on the task of detecting binary semantic change. Results in terms of precision, recall, and F1-score ("F1") are averaged across the 5 available words.

Ancient Greek	Precision	Recall	F1 score
GASC	0.600	1.0	**0.750**
SCAN	0.500	0.667	0.571
SGNS_NOT-technical	0.333	0.500	0.400
SGNS_NOT-narrative	0.333	0.500	0.400
SGNS_technical	0.000	0.000	0.000
SGNS_narrative	0.000	0.000	0.000
SGNS_NAIVE	0.333	0.500	0.400
TR_NOT-technical	0.400	1.0	0.571
TR_NOT-narrative	0.333	0.500	0.400
TR_technical	0.000	0.000	0.000
TR_narrative	0.500	1.0	0.667
TR_NAIVE	0.333	0.500	0.400

although a few steps in this direction have been taken (see e.g. Keersmaekers 2020, Rodda et al. 2019, and McGillivray et al. 2019). In spite of their limited use in lexical semantic change detection, dynamic Bayesian mixture models allow practitioners to embed domain expert knowledge and provide interpretable outputs.

We provided a systematic comparison of SCAN and GASC, two recent models from this family, with state-of-the-art embedding-based models, such as SGNS and Temporal Referencing. In addition, we transformed embedding models to account for genre information and provided a new evaluation framework to detect binary semantic change based on expert-annotated data.

Our experiments show that Bayesian models are highly competitive at detecting binary change, beating all baselines on Ancient Greek and Latin. These results, together with the ability to provide full representations of the evolution of word senses, indicate Bayesian dynamic mixture models as successful approaches to study semantic change in ancient languages.

This work can also be seen as a step towards the development of richer evaluation schemes and models that can embed expert judgement. We have shown how including genre can improve the understanding of the historical development of words in a corpus. We argue that the next process to be captured from semantic

change models is the archiving of historical texts. The entirety of the relevant documents extant at any time in history is an obvious reference population against which we perform inference. While any analysis based on a currently extant corpus could be biased, Bayesian models embedding historical domain knowledge enable us to de-bias the study (e.g., by accounting for missing texts when inferring the popularity of a sense). There are essentially three different necessary conditions for a text to be extant at any given time. The first is the decision of a librarian to add a particular document to a library, the second is whether or not that text is preserved or destroyed during the passage of time, and the third is the (in)ability of researchers to access documents extant at the current time. A Bayesian analysis enables us to embed a probabilistic description of such a development. For example, many texts within a given corpus will have their own associated provenance, which can be used to help inform the nature of the likely extant corpus. This allows historical insights and extra data to be drawn into the analysis and better inform historical conjectures. The explicit development of such models is ongoing, and we will report our findings in future work.

Acknowledgements

Abbreviations

NLP   natural language processing
SGNS   skip-gram with negative sampling

References

Ahmed, Amr & Eric P. Xing. 2012. Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. *UAI* abs/1203.3463.

Bamman, David & Gregory Crane. 2011. Measuring historical word sense variation. In *Proceedings of the 11th ACM/IEEE-CS Joint Conference on Digital Libraries*, 1–10. ACM.

Basile, Pierpaolo & Barbara McGillivray. 2018. Exploiting the web for semantic change detection. In *Discovery science* (Lecture Notes in Computer Science 11198). Springer-Verlag.

Biber, Douglas & Edward Finegan. 1989. Drift and the evolution of English style: A history of three genres. *Language* 65. 487–517.

Blei, David M. & John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning*, 113–120.

Boyd-Graber, Jordan, David Blei & Xiaojin Zhu. 2007. A topic model for word sense disambiguation. In *Proceedings of the 2007 EMNLP-CoNLL*, 1024–1033. http://www.aclweb.org/anthology/D/D07/D07-1109.

Cassotti, Pierluigi, Annalina Caputo, Marco Polignano & Pierpaolo Basile. 2020. GM-CTSC at SemEval-2020 task 1: Gaussian mixtures cross temporal similarity clustering. In *Proceedings of the 14th International Workshop on Semantic Evaluation*. Barcelona: ACL.

Cavallin, Karin. 2012. Automatic extraction of potential examples of semantic change using lexical sets. In *Proceedings of KONVENS 2012 (LThist 2012 workshop)*, 370–377. Vienna: ÖGAI.

Clackson, James. 2011. *A companion to the Latin language.* London: Wiley-Blackwell.

Clarke, Michael. 2010. Register Variation. In Egbert J. Bakker (ed.), *A companion to the Ancient Greek language*, 120–33. Chichester/Malden: Wiley-Blackwell.

Cook, Paul, Jey Han Lau, Diana McCarthy & Timothy Baldwin. 2014. Novel word-sense identification. In *Proceedings of COLING 2014: Technical papers*, 1624–1635. Dublin: ACL. https://www.aclweb.org/anthology/C14-1154.

Davies, Mark. 2002. *The corpus of historical American English (COHA): 400 million words, 1810-2009.* Provo: Brigham Young University.

Davies, Mark. 2012. Expanding horizons in historical linguistics with the 400-million word corpus of historical American English. *Corpora* 7(2). 121–157.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019: Volume 1 (Long and short papers)*, 4171–4186. Minneapolis: ACL. DOI: 10.18653/v1/N19-1423.

Dubey, Avinava, Ahmed Hefny, Sinead Williamson & Eric P. Xing. 2013. A non-parametric mixture model for topic modeling over time. *SDM* 2013. 530–538.

Dubossarsky, Haim, Simon Hengchen, Nina Tahmasebi & Dominik Schlechtweg. 2019. Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of ACL 2019*, 457–470. Florence: ACL. DOI: 10.18653/v1/P19-1044.

Dubossarsky, Haim, Daphna Weinshall & Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of EMNLP 2017*, 1136–1145. Copenhagen: ACL. DOI: 10.18653/v1/D17-1118.

Eger, Steffen & Alexander Mehler. 2016. On the linearity of semantic change: Investigating meaning variation via dynamic graph models. In *Proceedings of ACL 2016 (Volume 2: Short papers)*, 52–58. Berlin: ACL. DOI: 10.18653/v1/P16-2009.

Falk, Ingrid, Delphine Bernhard & Christophe Gérard. 2014. From the culinary to the political meaning of *quenelle*: Using topic models for identifying novel senses (De la quenelle culinaire á la quenelle politique : identification de changements sémantiques á l'aide des Topic Models). In *Proceedings of TALN 2014 (Volume 2: Short papers)*, 525–530. Marseille, France: Association pour le Traitement Automatique des Langues. https://www.aclweb.org/anthology/F14-2023.

Frermann, Lea. 2017. *Bayesian models of category acquisition and meaning development*. Edinburgh: University of Edinburgh. (Doctoral dissertation).

Frermann, Lea & Mirella Lapata. 2016. A Bayesian model of diachronic meaning change. *Transactions of the ACL* 4. 31–45. DOI: 10.1162/tacl_a_00081.

Frossard, Esteban, Mickael Coustaty, Antoine Doucet, Adam Jatowt & Simon Hengchen. 2020. Dataset for temporal analysis of English-French cognates. In *Proceedings of LREC 2020*, 855–859. Marseille: ELRA. https://www.aclweb.org/anthology/2020.lrec-1.107.

Frühwirth-Schnatter, S. 2006. *Finite mixture and Markov switching models*. 1st edn. Berlin: Springer.

Gale, William A., Kenneth W. Church & David Yarowsky. 1992. One sense per discourse. In *Speech and natural language: Proceedings of a workshop held at Harriman, New York, February 23–26, 1992*. https://www.aclweb.org/anthology/H92-1045.

Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of ACL 2016 (Volume 1: Long papers)*, 1489–1501. Berlin: ACL. DOI: 10.18653/v1/P16-1141.

Hengchen, Simon. 2017. *When does it mean? Detecting semantic change in historical texts*. Brussels: Université libre de Bruxelles. (Doctoral dissertation).

Hengchen, Simon, Nina Tahmasebi, Dominik Schlechtweg & Haim Dubossarsky. 2021. Challenges for computational lexical semantic change. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 341–372. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040322.

Jatowt, Adam, Ricardo Campos, Sourav S. Bhowmick, Nina Tahmasebi & Antoine Doucet. 2018. Every word has its history: Interactive exploration and Visualization of word sense evolution. In *Proceedings of CIKM 2018*, 1899–1902. ACM.

Jawahar, Ganesh & Djamé Seddah. 2019. Contextualized diachronic word representations. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 35–47. Florence: ACL.

Jenset, Gard B. & Barbara McGillivray. 2017. *Quantitative historical linguistics. A corpus framework*. Oxford: Oxford University Press.

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

Kessler, Brett, Geoffrey Numberg & Hinrich Schütze. 1997. Automatic detection of text genre. In *Proceedings of EACL 1997*, 32–38. ACL.

Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde & Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, 61–65. Baltimore: ACL. DOI: 10.3115/v1/W14-2517.

Korb, K. B. & A. E. Nicholson. 2009. *Bayesian artificial intelligence*. Cleveland, OH: CRC Press.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi & Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th international conference on the World Wide Web*, 625–635. Florence: ACM. DOI: 10.1145/2736277.2741627.

Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski & Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: A survey. In *Proceedings of COLING 2018*, 1384–1397. Santa Fe: ACL.

Kyrtatas, Dimitris J. 2007. Paradeisos. In Anastasios-Phoibos Christidis (ed.), *A history of ancient Greek: From the beginnings to Late Antiquity*, 1137–40. Cambridge: Cambridge University Press.

Lau, Jey Han, Paul Cook, Diana McCarthy, Spandana Gella & Timothy Baldwin. 2014. Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models. In *Proceedings of ACL 2014 (Volume 1: Long papers)*, vol. 1, 259–270. ACL.

Leiwo, Martti, Hilla Halla-aho & Marja Vierros. 2012. *Variation and change in Greek and Latin. Papers and monographs of the Finnish Institute at Athens*. Helsinki: Foundation of the Finnish Institute at Athens.

Liddell, Henry George, Robert Scott, Henry Stuart Jones & Roderick McKenzie. 1996. *A Greek-English lexicon*. 9th ed. Oxford: Oxford University Press.

McGillivray, Barbara. 2020. Semantic analysis of historical texts. In Kristen Schuster & Stuart Dunn (eds.), *Routledge international handbook of research methods in digital humanities*. London: Routledge.

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

McGillivray, Barbara & Adam Kilgarriff. 2013. Tools for historical corpus research, and a corpus of Latin. In Paul Bennett, Martin Durrell, Silke Scheible & Richard J. Whitt (eds.), *New methods in historical corpus linguistics*. Tübingen: Narr.

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

Mimno, David, Hanna Wallach & Andrew McCallum. 2008. *Gibbs sampling for logistic normal topic models with Graph-based priors*. Paper presented at the NIPS workshop on analyzing graphs, 2008, Whistler, BC.

Mitra, Sunny, Ritwik Mitra, Suman Kalyan Maity, Martin Riedl, Chris Biemann, Pawan Goyal & Animesh Mukherjee. 2015. An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering* 21(5). 773–798.

Mitra, Sunny, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee & Pawan Goyal. 2014. That's sick dude!: Automatic identification of word sense change across different timescales. In *Proceedings of ACL 2014 (Volume 1: Long papers)*, 1020–1029. Baltimore: ACL. DOI: 10.3115/v1/P14-1096.

Montariol, Syrielle & Alexandre Allauzen. 2020. Étude des variations sémantiques á travers plusieurs dimensions (Studying semantic variations through several dimensions). In *Actes de la 6e conférence conjointe journées d'études sur la parole (JEP, 33e édition), traitement automatique des langues naturelles (TALN, 27e édition), rencontre des étudiants chercheurs en informatique pour le traitement automatique des langues (réCITAL, 22e édition). Volume 2 : Traitement automatique des langues naturelles*, 314–322. Nancy: ATALA et AFCP. https://www.aclweb.org/anthology/2020.jeptalnrecital-taln.31.

O'Hagan, A. & J. Oakley. 2014. *SHELF: The Sheffield elicitation framework*. http://www.tonyohagan.co.uk/shelf/.

Pearl, J. 2009. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Dordrecht: Elsevier.

Perrone, Valerio, Paul A. Jenkins, Dario Spanò & Yee Whye Teh. 2017. Poisson random fields for dynamic feature models. *Journal of Machine Learning Research* 18(127). 1–45.

Perrone, Valerio, Marco Palma, Simon Hengchen, Alessandro Vatri, Jim Q. Smith & Barbara McGillivray. 2019. GASC: Genre-aware semantic change for Ancient Greek. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 56–66. Florence: ACL. DOI: 10.18653/v1/W19-4707.

Peters, Matthew, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee & Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT 2018: Volume 1 (Long papers)*, 2227–2237. New Orleans: ACL. DOI: 10.18653/v1/N18-1202.

Pollitt, Jerome Jordan. 1974. *The ancient view of Greek art: Criticism, history, and terminology* (Yale publications in the History of Art 26). New Haven: Yale University Press.

Rao, Vinayak & Yee Whye Teh. 2009. Spatial normalized gamma processes. In *Advances in neural information processing systems*, 1554–1562.

Rodda, Martina A., B. McGillivray & P. Probert. 2019. Vector space models of ancient Greek word meaning, and a case study on homer. *Traitement Automatique Des Langues* 60. 63–87.

Rodda, Martina A., Marco S.G. Senaldi & Alessandro Lenci. 2016. Panta rei: Tracking semantic change with distributional semantics in ancient Greek. In Pierpaolo Basile et al. (eds.), *Proceedings of third Italian conference on Computational Linguistics (CLiC-it 2016)* (CEUR Workshop Proceedings 1749). http://ceur-ws.org/Vol-1749/paper46.pdf.

Rosen-Zvi, Michal, Thomas Griffiths, Mark Steyvers & Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th conference on uncertainty in artificial intelligence*, 487–494.

Rudolph, Maja R. & David M. Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of WWW 2018*, 1003–1011. ACM. DOI: 10.1145/3178876.3185999.

Scarlini, Bianca, Tommaso Pasini & Roberto Navigli. 2020. Sense-annotated corpora for word sense disambiguation in multiple languages and domains. English. In *Proceedings of the 12th language resources and evaluation conference*, 5905–5911. Marseille, France: European Language Resources Association. https://www.aclweb.org/anthology/2020.lrec-1.723.

Schlechtweg, Dominik, Stefanie Eckmann, Enrico Santus, Sabine Schulte im Walde & Daniel Hole. 2017. German in flux: Detecting metaphoric change via word entropy. In *Proceedings of CoNLL 2017*, 354–367. Vancouver: ACL.

Schlechtweg, Dominik, Anna Hätty, Marco Del Tredici & Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic

change across times and domains. In *Proceedings of ACL 2019*, 732–746. Florence: ACL.

Schlechtweg, Dominik, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky & Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of SemEval 2020*, 1–23. Barcelona: ACL. https://www.aclweb.org/anthology/2020.semeval-1.1.

Schlechtweg, Dominik, Sabine Schulte im Walde & Stefanie Eckmann. 2018. Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of NAACL 2018*. ACL.

Schlechtweg, Dominik, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky & Barbara McGillivray. 2021. Dwug: A large resource of diachronic word usage graphs in four languages.

Shoemark, Philippa, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale & Barbara McGillivray. 2019. Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of EMNLP-IJCNLP 2019*, 66–76. Hong Kong: ACL.

Smith, Jim Q. 2010. *Bayesian decision analysis: Principles and practice*. Cambridge New York: Cambridge University Press, Cambridge.

Sprugnoli, Rachele, Marco Passarotti & Giovanni Moretti. 2019. *Vir* is to *moderatus* as *mulier* is to *intemperans*: Lemma embeddings for Latin. In *Proceedings of the Sixth Italian Conference on Computational Linguistics*. Torino: Accademia University Press. DOI: 10.5281/zenodo.3565572.

Tahmasebi, Nina. 2018. A study on Word2Vec on a historical Swedish newspaper corpus. In *Proceedings of the digital humanities in the Nordic countries, 3rd conference, Helsinki, March 7–9, 2018*. (CEUR Workshop Proceedings 2084). University of Helsinki, Faculty of Arts.

Tahmasebi, Nina, Lars Borin & Adam Jatowt. 2018. Survey of computational approaches to lexical semantic change. *arXiv preprint 1811.06278*.

Tahmasebi, Nina & Thomas Risse. 2017. Finding individual word sense changes and their delay in appearance. In *Proceedings of RANLP 2017*, 741–749. Varna: INCOMA Ltd. DOI: 10.26615/978-954-452-049-6_095.

Tang, Xuri, Weiguang Qu & Xiaohe Chen. 2016. Semantic change computation: A successive approach. *World Wide Web* 19(3). 375–415. DOI: 10.1007/s11280-014-0316-y.

Traugott, Elizabeth Closs & Richard B. Dasher. 2001. *Regularity in semantic change*. Cambridge: Cambridge University Press.

Vatri, Alessandro, Viivi Lähteenoja & Barbara McGillivray. 2019. *Ancient Greek semantic change: Annotated datasets and code [dataset]*. DOI: 10.6084/m9.Figshare.C.4445420.

Vatri, Alessandro & Barbara McGillivray. 2018. The Diorisis Ancient Greek corpus. *Research Data Journal for the Humanities and Social Sciences* 3(1). 55–65. DOI: 10.1163/24523666-01000013.

West, Mike & Jeff Harrison. 1997. *Bayesian forecasting and dynamic models.* 2nd edn. Berlin: Springer-Verlag.

Wijaya, Derry Tanti & Reyyan Yeniterzi. 2011. Understanding semantic change of words over centuries. In *Proceedings of DETECT 2011*, 35–40. ACM.

Wilkerson, Rachel L. & Jim Q. Smith. 2018. Customised structural elicitation. *arXiv preprint arXiv:1807.03693.*

Zamora-Reina, Frank D. & Felipe Bravo-Marquez. 2020. DCC-Uchile at SemEval-2020 Task 1: Temporal referencing word embeddings. In *Proceedings of the 14th International Workshop on Semantic Evaluation.* Barcelona: ACL.

Zampieri, Marcos, Shervin Malmasi & Mark Dras. 2016. Modeling language change in historical corpora: The case of Portuguese. In Nicoletta Calzolari et al. (eds.), *Proceedings of LREC 2016.* Portorož: ELRA.

Zhou, Jinan & Jiaxin Li. 2020. TemporalTeller at SemEval-2020 Task 1: Unsupervised lexical semantic change detection with temporal reference. In *Proceedings of the 14th International Workshop on Semantic Evaluation.* Barcelona: ACL.