Adjusting Scope: A Computational Approach to Case-Driven Research on Semantic Change

Lauren Fonteyn, Enrique Manjavacas

Leiden University Centre for Linguistics, Department of English Language and Culture, Arsenaalstraat 1, 2311CT, Leiden, the Netherlands

Abstract

Computational studies of semantic change are often wide in scope, aiming to capture and quantify semantic change in language at large in a data-driven, 'hands-off' way. Case-driven, corpus-linguistic studies of semantic change, by contrast, generally aim to tackle questions about the development of specific linguistic phenomena. Due to its narrower scope, case-driven research is more restricted in terms of the data it may employ, and at the same time it requires a more fine-grained description of the targeted linguistic developments. As a result, case-driven studies face particular methodological challenges that are not at play in more wide-scoped approaches. The aim of this paper is to set out a 'hands-off' computational procedure to study specific cases of semantic change. The case we address is the development of the phrasal expression *to death* from a literal, resultative phrase (e.g. *he was beaten to death*) into an intensifier (e.g. *We were just pleased to death to see her*). We deploy hierarchical clustering algorithms over distributed meaning representations in order to capture the evolution of the semantic space of verbs that collocate with *to death*. We then describe the arising diachronic processes by means of monotonic effects, providing a more accurate picture than customary linear regression models. The methodology we outline may help tackle some common challenges in the use of vector representations to study similar cases of semantic change. We end the discussion by pinpointing (remaining) challenges that case-driven research may encounter.

Keywords

Linguistics, Semantic Change, Grammaticalization, Distributional Semantics, Bayesian Modeling

1. Introduction

Over the past decade, computational approaches to semantic change have experienced a surge in popularity. This is largely due to the rise of an increasingly powerful body of models that aim to approximate the meaning of words over time by encoding their linguistic context (or 'distributional properties') into (diachronic) word embeddings [see, among many others: 47, 19, 35, 18, 1, 44, 46, 29, 26, 13, 51, 11, 48, 5, 50]. A characteristic of many of these studies is that their research questions are very wide in scope: their aim is not to address questions about any specific word or construction, but rather to capture and quantify some aspect of semantic change at large in a data-driven way. As such, these studies tend to approach semantic change in bulk, with sample sizes ranging from hundreds [e.g. 35, 13, 48] to thousands of linguistic items [e.g. 18], and with specific examples of semantic change predominantly serving as (straightforward) illustrations of a more general pattern or trend.

CHR 2021: Computational Humanities Research Conference, November 17–19, 2021, Amsterdam, The Netherlands

l.fonteyn@hum.leidenuniv.nl (L. Fonteyn); enrique.manjavacas@gmail.com (E. Manjavacas)
0000-0001-5706-8418 (L. Fonteyn); 0000-0002-3942-7680 (E. Manjavacas)

Yet, vector-based models are also used in more narrow-scoped, case-driven research. In this type of research, which is perhaps most common in (corpus) linguistics, the aim is to approach a specific case study of semantic change in a largely automated, data-driven manner. The motivation for doing so is that introspective data annotation (which is not only labour-intensive, but potentially problematically subjective [47]) is avoided or minimized. Additionally, the use of vector-based models allows researchers to operationalize theoretical concepts in quantifiable terms in order to verify or falsify hypotheses on the nature and causes of semantic change in the case under scrutiny [e.g. 21, 10, 41, 40].

Despite the fact that the two approaches have an obvious common ground, case-driven investigations are clearly distinct from wide-scope computational studies in a number of ways. Most importantly, case-driven research generally emerges from a desire to tackle questions about the development of specific linguistic phenomenon (often during a specific time window). Consequently, compared to the wide-scope computational research into semantic change, case-driven research is relatively inflexible in terms of the data it may employ to attain its goals. Furthermore, the very reason why the researcher is compelled to undertake case-driven research is that the specific phenomenon under scrutiny constitutes a complex challenge. As such, while it is not uncommon for case-driven studies to use computational models and metrics proposed in wide-scoped studies on semantic change, the specific cases they scrutinize may present methodological challenges that are either not at play, or glossed over in, the studies they draw from.

The aim of the present contribution is to tackle a case-driven study by means of computational methods. In doing so, our work ties in with earlier explorative work aimed at pinpointing where challenges may lie for case-driven research [e.g. 49]. The specific case of semantic change we address is the historical development of the phrasal expression *to death* from a literal, resultative phrase (e.g. *he was beaten to death*) into an intensifier (e.g. *We were bored/pleased to death*) [23, 33].

1.1. Aims

More specifically, we aim to delineate a step-wise procedure that:

1. minimizes manual work, so that it is more feasible for case-driven research to maximally exploit available data and maximize the data-driven character of case-driven research;
2. flags and discusses remaining pitfalls and challenges future case-driven work may encounter.

1.2. Outline

To analyse the development of *to death* with minimal manual interference, we suggest a procedure consisting of the following steps:

1. Surveying work on the linguistic construction (and related cases) under scrutiny to (i) delineate a time window, and (ii) formulate hypotheses or expectations that can be verified by means of computational methods (Section 2);
2. Compiling (and curating) a sufficiently large diachronic corpus collection (Section 3), from which examples of the construction can be sampled (Section 3.1);
3. Computing and evaluating distributed meaning representations (Section 3.2);

4. Conducting a diachronic cluster analysis, in which we optimize the number of clusters across time for silhouette score in order to trace changes in *to death*'s contextual distribution (Section 4.1);
5. Conducting a sentiment analysis to capture *to death*'s decreasing negativity (Section 4.2);
6. Assessing the output of the statistical model against the formulated expectations (Section 5).

After describing the procedure and results, we highlight and discuss the following remaining pitfalls (Section 6):

1. Because case-driven research aims to examine a fixed (set of) linguistic construction(s) in a specific time window, researchers may run into issues of data sparsity and balance that may be difficult to circumvent.
2. The extent to which a reliable, completely automated, 'hands-off' approach is possible and extendable to new cases not yet analysed remains an open question. While we are optimistic about incorporating computational models into the study of case-driven semantic change, manual interference may still be desirable or even required.

2. Related Work

The methods adopted in the present study are similar to those used in large-scale computational studies of semantic change. Previous work has suggested various ways of improving the models that generate (diachronic) word embeddings [e.g. 46, 44], determining (predictive) laws of (lexical) semantic change at large [e.g. 19, 14], and developing statistical measures that help detect different types of semantic change (e.g. specification vs. broadening; cultural change vs. linguistic change) in a data-driven manner [e.g. 47, 35, 18, 11, 13, 48, 16]. In other work, computational models are applied to map changes in specific word classes (e.g. intensifiers; [31], or (groups of) concepts in particular lexical domains (e.g. 'racism', 'knowledge'; [49, 2]) or registers (e.g. 'scientific language'; [5, 50]).

In terms of its focus, aims, scope and granularity, this study is reminiscent of research in corpus linguistics and construction grammar, where a single case of linguistic change is considered. The development of *to death* from a phrase that expresses the result of an action (e.g. *He was beaten/stabbed/shot to death*) to an intensifying or 'amplifying' expression (e.g. *We were thrilled/pleased/shocked to death to see you*; [42]) has been described as a process of grammaticalization, which took place over the course of the Early and Late Modern English period (ca. 1500 - Present). As explained by Margerie [33], this grammaticalization process, in which *to death* developed a less literal and more 'grammatical' reading of amplification, crucially involved 'host-class expansion' [22]. More specifically, the development can be broken down into three stages.

STAGE 1 Initially, *to death* functioned as adverbial complement of verbs expressing physical harm, which may result in death (e.g. *beat*, *bleed*, *burn*, etc.).

STAGE 2 Over the course of the 16th and 17th century, *to death* sporadically started occurring in contexts where a literal, death-resulting reading is ruled out (e.g. *That book bored me to death*). It was not until the 18th century, however, that *to death* was frequently used in such non-literal, intensifying cases [33, 23]. Notably, as is common in intermediate stages of grammaticalization [25, 30], *to death* still retained some of its original meaning of a 'negative end result' [33, p. 129]. At this stage, the vast majority of its collocate verbs have negative connotations (e.g. *bore, scare, worry*).

STAGE 3 Despite its persistent preferences for negative situations, *to death* started to expand further [33]. In the 19th and 20th century, *to death* began to combine with more positively oriented verbs (e.g. *amuse, love, thrill*).

The expansion process spawned by the grammaticalization of *to death* would seem to lend itself well to computational analysis. A template for the general research design can be found in the work of Perek [41, 40]. With an eye on quantifying processes related to host class expansion, Perek relies on semantic vector representations of the verb types occurring in the constructions open verb slot of the *hell*-construction (e.g. *[beat/scare/hug] the hell out of someone*) and the *way*-construction (e.g. *[swim/beat/smile] one's way to something*), employing cluster density measures in order to quantify the diachronic process.

Crucially, Perek demonstrates that, from a linguistic perspective, it is important to approach processes of host class expansion in a way that distinguishes changes in lexical diversity (measured by the number of unique lexical items that occur in a construction) from semantic diversity (measured by the semantic similarity between those lexical items). This is also relevant for the study of *to death*, because changes in lexical diversity alone may not be indicative of linguistic change, but of cultural change. It may be the case, for instance, that different modes of execution have become prevalent or obsolete, or that the specificity and lexical diversity with which causes of death are described may increase or decrease as the topic becomes more or less taboo. In these scenarios, the set of lexical items *to death* collocates with may indeed shrink or expand, but the semantics of the phrase do remain stable. At the same time, such cultural change may happen alongside the grammaticalization of *to death* into an intensifier. Thus, the reality of case-driven research may be that the distinction between cultural and linguistic change is not a matter of 'either/or' [18], but of 'and'.

The distributed meaning representations that are fed into the clustering algorithm by Perek [41, 40] do, however, fail to distinguish synonymy (e.g. *hate* & *despise*) from antonymy (e.g. *hate* & *love*). Hence, they will not capture the the final stage of expansion of *to death* and other intensifying constructions, which commonly involves an erosion of its original negative (or positive) polarity [30].

3. Data

For the purposes of the present study, we gathered a collection of diachronic English corpora, spanning the period from 1550 to 1949. These corpora include Early English Books Online (EEBO), the Corpus of Late Modern English Texts (version 3.1; CLMET3.1), the Evans Early American Imprints Collection (EVANS), Eighteenth Century Collections Online (ECCO), the Corpus of Historical American English (COHA), and the Hansard corpus (Hansard). In terms of text types, these corpora are varied, covering an array of literary works, religious and legal text and news reports. The sole exception is Hansard, which offers transcriptions of British parliamentary debates (starting in 1800).

All corpora were submitted to the following pre-processing pipeline. First, we applied a language identification module in order to sort out foreign text. We relied on two language identification modules – Google's Compact Language Identifier (v3)[1] and FastText Language

[1] Our code repository is accessible through the following url: https://github.com/google/cld3/releases/tag/3.0.13.

Table 1
Distribution of *to death* per bin (by corpus) and verb type frequency in the sample (last row).

	1550	**1600**	**1650**	**1700**	**1750**	**1800**	**1850**	**1900**	***Total***
CLMET3.1				39	45	182	100	26	***392***
COHA						488	700	774	***2764***
ECCO				78	395	12			***485***
EEBO	800	800	794	413		2			***2859***
EVANS			6	211	360	116			***693***
Total	***800***	***800***	***800***	***741***	***800***	***800***	***800***	***800***	7193
Type Freq	87	101	93	95	97	150	131	135	***372***

Identification system [17] – which we combined to maximize the retrieval precision of the foreign text. For a given fragment of 500 characters, we flagged the text as foreign if both systems indicated a language other than English as the highest probability language. Manual inspection of a random sampled indicated a sufficiently low false positive rate in order for the filtering to be effective (while throwing out an insignificant amount of English text).

Second, we tokenized and sentence-tokenized the remaining text using the Punkt tokenizers provided by the NLTK package [4]. After tokenization, we enriched all text with part-of-speech tags, using an in-house tagger for historical English. The tagger was trained on the PCEEME [36] – a corpus of letters from 1410 to 1695 that amounts to about 2.2M labelled tokens – using a Neural Conditional Random Field (CRF) tagger implemented with PIE [32], and obtained an overall test-set accuracy above 96%.

The resulting patchwork corpus consists of a total of 3.9B tokens, which we utilized in various ways in subsequent steps of the research process.

3.1. Dataset: *to death*

The attestations of *to death* were retrieved from the corpus collection (excepting the specialized Hansard corpus). As is common in linguistic research, the data was divided into fixed-width bins. Each bin represents a 50-year period, which results in a total of 8 bins. As not all corpora in the collection are balanced in terms of the amount of text a single author may contribute, we applied an additional sampling step to ensure that no author dominated more than 25% of the instances in a particular bin.

The total number of instances retrieved from each corpus per bin is listed in Table 1. In the bin covering the period between 1700 and 1749, the total corpus size (and hence, the token frequency of *to death*) was substantially lower than for other bins. To ensure that any observed differences in the number of verb types that collocate with *to death* across bins is not affected by large differences in sample size, we decided to cap the maximum number of tokens sampled per bin at 800.

After removing any duplicates, we identified the verb that collocates with each instance of *to death* by relying on part-of-speech tags. Each instance of *to death* was assumed to collocate with the verb in closest proximity (using a window of 15 words). In a number of cases, the tagger failed to find a collocate verb. These cases included instances where the copula *be* was used in combination with an adjective (e.g. *be frozen/sick to death*), which were subsequently corrected and included in the dataset. Cases where *to death* functioned as a prepositional modifier of a noun (e.g. *on her way to death*), fixed expressions (e.g. *from birth to death*, *be*

Table 2
Word embedding benchmark results for the utilized word embedding space in comparison to off-the-shelf Present-day English spaces.

	MEN	WS353	SimLex999	MTurk	RW	RG65	Mean
Glove	0.608	0.399	0.331	0.513	0.283	0.736	0.478
Word2Vec	**0.708**	**0.605**	**0.414**	**0.645**	**0.378**	**0.746**	**0.583**
Ours	0.555	0.504	0.338	0.481	0.241	0.731	0.475

nigh to death), and cases where the verb was illegible (e.g. *And when my mother euen before my sighte, Was (-) to death*; 1550, EEBO) were discarded. In total, 109 examples were discarded.

3.2. Word Embeddings

In order to capture semantic similarity between *to death*'s collocate verbs across time, we rely on distributed meaning representations computed by the `word2vec` algorithm [34]. We use the entire corpus collection introduced in Section 3. Besides the pre-processing pipeline outlined in Section 3, we applied the following additional pre-processing steps with the goal of improving the quality of the resulting embeddings: we lower-cased the corpora, applied NFKD unicode normalization, removed non-alphanumeric tokens, replaced numbers by a code (e.g. `<NUM>`), dropped punctuation, and substituted the long "s" character (f) with modern day "s".

We trained distributed representations with a size of 200 using the `gensim` library [43]. We employed the skip-gram objective, approximated with negative sampling and optimized using a learning rate of 0.025 over 5 epochs, discarding words with frequencies lower than 50 and a window size of 20 tokens.

In order to validate the resulting embedding space, we ran a number of semantic similarity benchmarks, which allow us to contextualize the quality of our embeddings within the state-of-the-art. The employed benchmark datasets comprise of sets of Present-day English word pairs, each of which has been manually assigned a similarity score. The evaluation proceeds by correlating these human judgments with the cosine similarities between the corresponding vector representations, using the Spearman correlation coefficient.[2] We compared our embedding space with (i) 200 dimensions `Glove` vectors [39] trained on 6B Wikipedia tokens,[3] as well as (i) 300 dimensions `word2vec` vectors trained on the Google News dataset (about 100B tokens), restricting the vocabulary of the embedding spaces to the intersection across spaces and using the average word embedding vector for out-of-vocabulary words.[4]

As Table 2 shows, our embedding space generates scores comparable to the `Glove` space, while lying behind those generated by the `word2vec` space. Considering that our embedding space is trained on a smaller dataset and covers a large period of historical English, we take these results to validate the semantic similarity properties of the inferred word representations. For a sanity check, Table 3 shows the 20 nearest neighbours of a selection of verbs from our dataset of *to death* collocates based on cosine distance.

[2] While it is obviously not ideal to evaluate our model with respect to a Present-day English reference point, no human similarity judgements of this scale are available for historical English. In order to conduct at least some sort of sanity check, we used the off-the-shelf Present-day English spaces.

[3] The embeddings are available through the following url: https://nlp.stanford.edu/projects/glove/.

[4] We use the software package `word-embedding-benchmarks` [27] in order to streamline the evaluation of the embedding spaces.

Table 3
Top 10 nearest neighbours (cosine) of *burn*, *stab*, *whip* (physical actions) and *amuse*, *scare*, *vex* (mental verbs) in collocate dataset.

	physical actions			*mental verbs*		
	burn	**stab**	**whip**	**amuse**	**scare**	**vex**
1	beat (0.57)	strangle (0.59)	cudgel (0.69)	delude (0.73)	frighten (0.78)	afflict (0.72)
2	kill (0.57)	knife (0.59)	bludgeon (0.66)	flatter (0.63)	terrify (0.73)	perplex (0.72)
3	consume (0.56)	bleed (0.58)	lash (0.66)	perplex (0.61)	startle (0.67)	harass (0.71)
4	scorch (0.55)	slash (0.58)	kick (0.59)	terrify (0.60)	worry (0.55)	annoy (0.69)
5	shoot (0.55)	bang (0.56)	cuff (0.57)	frighten (0.60)	drive (0.54)	oppress (0.69)
6	spoil (0.53)	kill (0.55)	spur (0.57)	tickle (0.58)	sweep (0.52)	fret (0.67)
7	smother (0.53)	poison (0.55)	flog (0.56)	harass (0.54)	delude (0.51)	grieve (0.64)
8	smoke (0.53)	bite (0.55)	bang (0.55)	tire (0.54)	astonish (0.51)	terrify (0.61)
9	hunt (0.53)	cudgel (0.54)	goad (0.55)	annoy (0.52)	annoy (0.50)	pester (0.60)
10	hang (0.53)	prick (0.54)	scourge (0.54)	vex (0.51)	amuse (0.50)	worry (0.58)

4. Method

A basic way of quantifying the host class expansion of *to death* is by examining the change in diversity in the set of attested collocate verbs over time. One such index of diversity is given by type frequency – shown in the last row of Table 1. However, while such diversification is potentially indicative of host class expansion, changes in type frequencies (or lexical diversity) need not indicate that *to death* has indeed undergone semantic change; as argued in Section 2, they may equally be indicative of cultural change. To probe into the host class expansion of *to death*, Section 4.1 operationalizes the process as a change in the structure of the semantic space that the collocate verbs of *to death* occupy. We rely on hierarchical cluster analysis over distributed meaning representation in order to not only incorporate a notion of lexical diversity into the analysis but, crucially, also take ‘semantic diversity’ into account.

As explained in Section 2, the host class expansion of *to death* also involved increased co-occurrence with verbs with progressively more positive connotations. In order to capture this process, Section 4.2 devises a way to quantify the average polarity of verbs over time using word embeddings, and statistically describe any existing ‘positivization’ process.

4.1. Cluster Analysis

At any given period, we inspect the semantic space delineated by the distribution of attested verbs using a hierarchical cluster analysis. A known problem with automated cluster analysis of semantic spaces is that the induced semantic clusters are not always easy to interpret. As a result, their application in subsequent steps of the research workflow may require manual fine-tuning and post-filtering [41, 40] to ensure that clusters are meaningful before any measures of interest can be computed. In contrast, the chosen procedure dispenses with manual fine-tuning and inspection of the resulting clusters. First, we identify a clustering metric that aligns with the expectations of the host-class expansion process. Secondly, we automatically find the hyper-parameter values that optimize the selected clustering metric, and, finally, we treat these optimal values as statistical correlates of the host-class expansion process that we are ultimately interested in describing.

As *to death* develops new, non-literal meanings, we expect the semantic space defined by the verbs appearing in this construction to expand, with existing clusters of collocates becoming denser and new clusters representing novel semantic fields starting to form. The silhouette score [45] – a common clustering evaluation metric – may help determine whether this expectation holds up. More specifically, we use the *optimal number of clusters based on the silhouette score* as the target statistic for monitoring the process.

For a word w_i assigned to cluster C_i, the silhouette score decomposes into the quantities $a(w_i)$ – shown in Equation 1 – and $b(w_i)$ – shown in Equation 2. By measuring the average intra-cluster distance between a word and all other words in the same cluster, $a(w_i)$ captures the *tightness* of the clusters induced by a clustering algorithm. In contrast, $b(w_i)$ measures the distance to the nearest point in a different cluster. The dataset-level aggregated $b(w_i)$, thus, captures the overall *separation* between clusters.

$$a(w_i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} \cos_{dist}(w_i, w_j) \tag{1}$$

$$b(w_i) = \min_{k \neq i} \sum_{j \in C_k} \cos_{dist}(w_i, w_j) \tag{2}$$

The final silhouette score for a given instance is computed by an aggregation of both quantities, dividing by a normalizing factor to ensure a constant output range between -1 and 1 – as shown in Equation 3.

$$s(w_i) = \frac{b(w_i) - a(w_i)}{\max(a(w_i), b(w_i))} \tag{3}$$

One risk that can be linked to the presented methodology is that the optimal number of clusters may increase because the number of unique verb types in the sample has increased (as shown in Section 3) – i.e. regardless of the semantic composition of the space representing that bin. Thus, increases in the optimal number of clusters can be due to sampling artifacts – an issue that becomes even more likely with fat-tailed distributions that are common in linguistic data. Moreover, even in the absence of sampling artifacts, we must ensure that we are not simply measuring increases in type frequency-based diversity, which, as already argued, are not necessarily indicative of linguistic change.

In order to remedy the afore-mentioned issue, we employ the following bootstrap procedure. For each period, we sample 500 verbs with replacement from the multinomial distribution observed in the dataset and compute the optimal number of clusters based on silhouette score. Repeating this process a 1,000 times per period yields a dataset with 8,000 observations (i.e. for 8 periods), which we submit to statistical analysis in order to quantify the effect of time on the optimal number of clusters. Crucially, we record the total number of *distinct verbs* sampled in each bootstrap iteration, which allows us to statistically control for the effect of population size on the obtained optimal number of clusters.

We rely on hierarchical (agglomerative) clustering using the cosine similarity and complete linkage,[5] and optimize the number of clusters by inspecting the silhouette scores at different nodes in the induced merge tree until reaching the merge step that maximizes the silhouette score.[6]

[5] We made these choices on the basis of a single manual scan of the interpretability of the clusters induced from the verbs in the entire dataset.

[6] We use the reference implementations provided by the Python library `scikit-learn` [38].

4.2. Sentiment Analysis

Homing in on the increasing positivity of *to death*, we leverage the embedding space described in Section 3.2 in order to capture the sentiment polarity of the sampled verbs. Differences in sentiments are not straightforwardly captured by means of hierarchical clustering, as antonyms are represented by highly similar vectors. In Table 3, for instance, the positive mental verb *amuse*, is recognized as being similar to more negative mental verbs like *delude* and *terrify*, as well as its antonyms *annoy* and *vex*. The cluster analysis is therefore supplemented by means of sentiment scores.

A first approach to induce word-level sentiment scores is to exploit the proximity of a given verb vector to the vector for the words ‘good’ and ‘bad’. The closer to the vector for ‘good’ the more positive the sentiment of that verb. However, similar confounding effects from antonyms make this approach unfeasible. Indeed, in common word embedding spaces the vectors for ‘good’ and ‘bad’ tend to be located in the proximity of each other, and, thus, lack discriminative power for classifying words with respect to their sentiment.

In order to tackle this issue, post-hoc modifications of the embedding space such as retrofitting [15] or word embedding refinement [53] could allow us to leverage sentiment lexicons in order to ensure the desired property. In the present work, however, we dispense with the manual work that such approach would require and resort to a second-order approach that induces sentiment scores on the basis of the proximity of verbs to a filtered list of nearest neighbors of ‘good’ and ‘bad’. By manually filtering these lists, we avoid terms that may confound the polarities, while still keeping the manual work to a small amount. More specifically, we sift through the vocabulary in ranked order by cosine similarity to “good” and “bad”, and discard confounding words until reaching a total of 20 words per polarity.[7] For a given word w_i, we, then, compute its sentiment score as shown in Equation 4:

$$S(w_i) = \frac{1}{|N_{good}|} \sum_{w_j \in N_{good}} \cos(w_i, w_j) - \frac{1}{|N_{bad}|} \sum_{w_j \in N_{bad}} \cos(w_i, w_j) \tag{4}$$

where N_{good} and N_{bad} refer, respectively, to the filtered set of nearest neighbours of ‘good’ and ‘bad’.

To test the effect of time on the polarity of *to death*’s collocates, we assign each verb in the dataset to the bin where they are first attested. Given that grammaticalizing structures often retain their original function, it may well be that the well-established negative use of *to death* vastly outnumbers and hence overshadows cases where *to death* has expanded to intensify new, more positive verbs. Thus, we suggest that working with the sentiment of collocate verbs that were first attested in a given bin – rather than the distribution of sentiment in each bin – captures the ongoing changes more directly and robustly.

4.3. Statistical Modeling

In order to assess the effect of time on the semantic structure of the attested verbs, as well as on the overall sentiment, we fit linear regression models regressing the target outcome – i.e. optimal number of clusters or sentiment score – on the time period. We use a Gaussian likelihood for both outcomes.

[7] These filtered nearest neighbors were checked in order to avoid too specific terms with unstable sentiment polarity over time. For example, the top 5 neighbours of ‘good’ were ‘better’, ‘excellent’, ‘great’, ‘well’ and ‘best’, while the top 5 neighbours of ‘bad’ were ‘dangerous’, “ill”, ‘inefficient’, ‘wrong’ and ‘hard’.

A further modeling choice we make is to incorporate time period as a monotonic effect – and not as, for instance, an ordinary linear predictor. This choice is motivated by the fact that diachronic processes in language structure often result in patterns that resemble s-curves [12, 6]. In these patterns, the magnitude of the predictor varies over time, a fact that cannot be described by ordinary linear predictors. In contrast, a monotonic predictor shares the assumption with a linear predictor that the direction of the effect is constant – strictly positive or negative – while allowing differences in the effect over adjacent time periods.

Our implementation of the monotonic predictor follows Bürkner and Charpentier [8]. For a given predictor with n possible categories (in our case, this corresponds to 8 time bins) to be modelled as a monotonic effect, this approach introduces n-1 ζ_i parameters such that $\zeta_i \in [0, 1]$ and $\sum_{i=1}^{n-1} \zeta_i = 1$, keeping ζ_0 fixed at 0. For a given observation of the j^{th} time bin, the monotonic predictor term η is given by Equation 5:

$$\eta = b \sum_{i=1}^{j} \zeta_i \tag{5}$$

Here, b corresponds to the ordinary linear predictor, representing in this case the direction and size of the effect on the outcome, and the individual ζ_i represent the normalized distances between consecutive predictor categories. The predictor term η is then included in a linear model in the usual way: $y = a + \eta \times x$. When fitted, this kind of monotonic predictor can easily be interpreted by inspecting the values assigned to the ζ_i parameters, since these correspond to the relative increase of each category with respect to the total increase involved by the monotonic predictor.

5. Results

We deploy a Bayesian regression framework which allows us to inspect the uncertainty in the statistical parameters of interest in an probabilistic intuitive manner. We fit our models using the Hamiltonian Monte-Carlo sampler provided by the `stan` library [9] through the **R** language package `brms` [7].

5.1. Cluster Analysis

In order to test the monotonicity of the effect, we compare a linear model of the effect of time period on the optimal number of clusters – `LINEAR(P)` – with the monotonic effect model – `MONO(P)`. Moreover, in order to control for the size of the sampled population on the outcome, we fit additional models including the number of unique verbs in the bootstrap sample as predictor – `LINEAR(P)+S` and `MONO(P)+S`.

We compare the four models using the Widely Applicable Information Criterion (WAIC), which estimates the plausibility of the models in terms of both predictive performance and model complexity (cf. overfitting). The results of the comparison are shown in the top row of Table 4. Including time period as a monotonic effect improves the predictive power of the model over the linear effect. Moreover, controlling for sample size is even more important, as evidenced by the fact that including it results in a larger improvement in WAIC than modeling period as a monotonic effect.

Using the most strongly predictive model – i.e. `MONO(P)+S` – we can visualize the (monotonic) effect of time period on optimal number of clusters using the posterior predictive distribution.

Table 4

Comparison of statistical models of optimal number of clusters and polarity using the WAIC criterion (lower is better). Besides absolute WAIC, we also show an estimate of the effective number of parameters (P), the difference in WAIC (WAICΔ(SE)) and the model weight (Weight), quantifying the relative value of each model with respect to the remaining models.

Outcome	Model	WAIC (SE)	P	WAICΔ(SE)	Weight
Clusters	`MONO(P)+S`	50,777 (137)	5.73		1.00
	`LINEAR(P)+S`	50,995 (136)	4.30	-218.42 (25.10)	0.00
	`MONO(P)`	52,211 (132)	4.86	-1,434.14 (71.02)	0.00
	`LINEAR(P)`	55,780 (128)	2.75	-5,003.14 (117.48)	0.00
Polarity	`MONO(P)`	784.82 (25.08)	3.73		0.79
	`LINEAR(P)`	787.52 (25.03)	3.18	-2.7 (1.38)	0.21

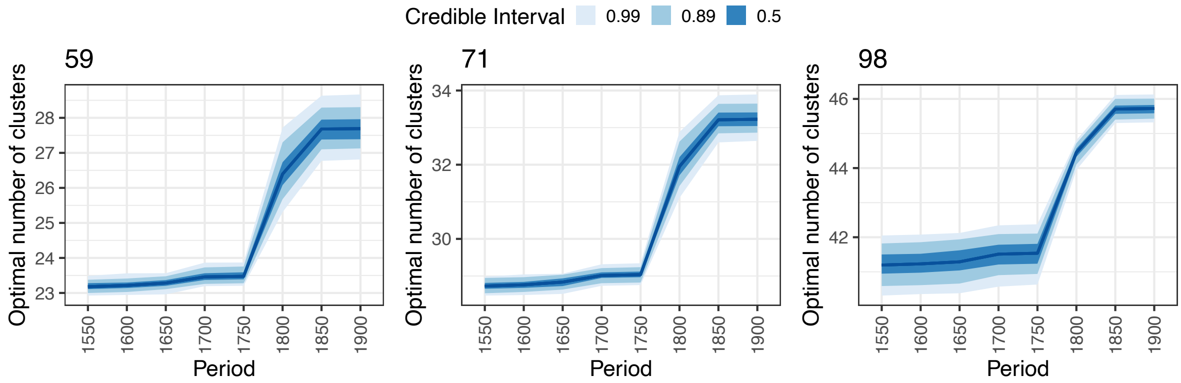

Figure 1: Posterior predictive distribution of the optimal number of clusters by period, showing different credible intervals, while varying the sample size over 59 (left), 71 (middle) and 98 (right) items, corresponding respectively to the 10%, 50% and 90% percentiles. The visualization is based on 200 samples from the MCMC posterior draws.

Figure 1 depicts the posterior predictive distribution of the optimal number of clusters using a counter-factual triptych plot, statistically controlling for the sample size at different percentiles. Overall, we observe a clear monotonic effect, resembling an s-curve, with a leap starting in the 1750 bin. The shape of the effect remains stable across the three sample size percentiles. Due to the positive linear effect of sample size on optimal number of clusters, the range of the outcome (i.e. the y-axis) increases across plots in the triptych. Moreover, the distribution of uncertainty varies from plot to plot. At smaller sample sizes, the uncertainty in the predicted number of clusters is larger towards the later time bins, whereas for larger sample sizes the most uncertain predictions come from the earlier bins. This is likely due to the fact – depicted in Figure 2 – that the sample size in pre-1800 bins is always smaller than in post-1800 bins. However, by counter-factually controlling for sample size, we can observe that the statistical model predicts a constant effect shape regardless of the sample size.

5.2. Sentiment Analysis

Similarly to the experiments in Section 5.1, we now compare the effect of time period on sentiment using a linear predictor – `LINEAR(P)` – and a monotonic effect – `MONO(P)`. We use

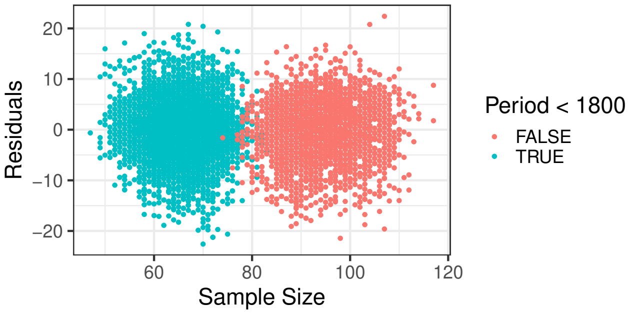

Figure 2: Residuals of the model by sample size. Color highlights are used to distinguish pre-1800 and post-1800 observations. Despite the increase in sample size starting in 1800, residuals do not seem to be correlated with sample size.

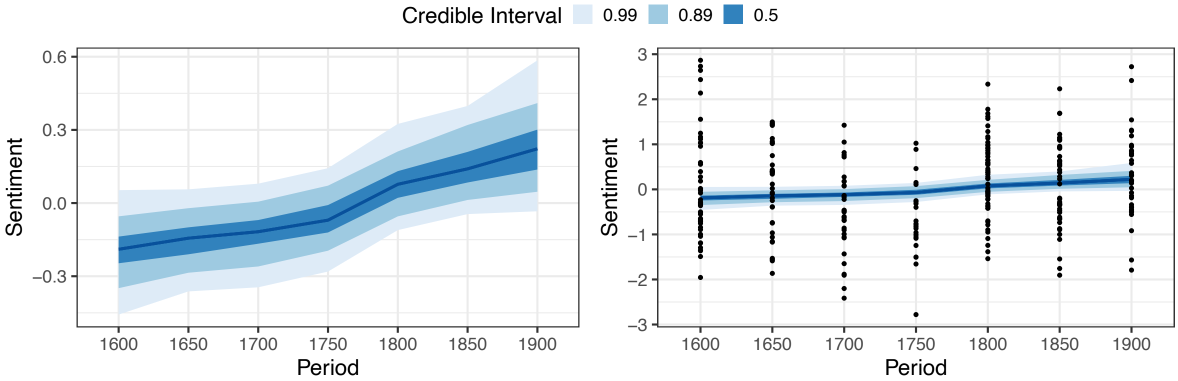

Figure 3: Posterior predictive distribution of the statistical model of sentiment using time period as monotonic effect (left), posterior predictive distribution with overlaid empirical observations (right).

the standardized average sentiment polarity of the verbs as the outcome. The results in terms of WAIC are shown in the bottom row of Table 4. Modeling time with a monotonic effect produces an improvement over the linear predictor, although in this case the difference with respect to the linear effect model is smaller than in the cluster analysis experiments. The left plot shown in Figure 3 does indicate a slight jump starting in the 1750 bin. However, the large credible intervals observed do not rule out a merely linear effect. Moreover, as the plot in the right hand-side of Figure 3 shows, a considerable amount of variance in the dataset is left unexplained by the model. While statistically controlling for other predictors – such as, for example, document topic or genre – could improve the fit, the current model does show a predominantly linear upward effect of moderate size – about 1 standard deviation – of time on average sentiment.

6. Discussion

The results of the statistical analyses are in line with expectations in that the optimal number of verb clusters increases substantially over the course of the 18th century, when the meaning of *to death* expanded to non-literal, intensifying uses (STAGE 2). The predicted shift away from negative polarity (STAGE 3) also appears to be captured by the statistical model, albeit weakly. Still, as even in present-day English *to death* is predominantly attested with negative collocates [33], the weak trend aligns well with the pathway outlined in Section 2. All in all, then, the procedure adopted here is promising for future case-driven, 'hands-off' investigations.

With an eye on aiding future applications of the models and methods adopted in the present study, we highlight some important remaining problems.

6.1. Data sparsity and balance

§While there is no shortage of Historical English corpora, corpora that span all the way from the Early Modern period up to Present-day English are rare. A notable exception is the suite of the Penn-Helsinki Corpora [28], which, although wide in scope, is still a corpus collection very limited in size, and thus also in its use for the 'data-hungry' models that are currently employed in computational studies of semantic change.[8] To maximize sample sizes, this study (following Margerie [33]) resorted to combining large corpora covering different time windows.

An issue with this patchwork solution is that individual time bins are likely not represented by a comparable number of texts and text types, which may have consequences at later steps in the procedure. In the present case, the patchwork corpus suffered from data sparsity in the 1700-1749 bin, which in turn forced us to cap the maximum number of tokens per bin. Furthermore, because of the inconsistency with which text types are labelled across the different corpora, it is very difficult if not impossible to smoothly ensure register and genre consistency across bins. For the present case, such text type inconsistency is indeed very unfortunate: the time bin in which the host class expansion of *to death* has taken off also appears to be the time bin in which the COHA corpus starts, which introduces newspaper and magazine texts into the sample.[9] At the same time, some of the text collections included in the patchwork corpus (such as ECCO) may contain reprints of older texts, which may have led to an overrepresentation of older usages in certain time bins. As such, a limitation of the procedure presented here is that it devotes relatively limited attention to balancing data and/or controlling for text and text type variation across time bins. A possible solution could be to refrain from working with corpus patchworks, and turn to the Google Books Corpus (1500 - 2008) or other large library dumps. Yet, even then issues of overrepresentation (and mislabelling) of texts and text types may remain [52, 37].[10]

In short, a substantial challenge for case-driven research is that the careful data curation it requires may lead to an impasse. Even when following strict procedural guidelines and sanity checks [49, 14], artefact results are still possible when data is uncurated or poorly balanced (as also discussed for lexical semantic change in Hengchen et al. [20]). Furthermore, introducing such balance may not be easy (or even possible), and it may also impact sample size (which complicates the study of more infrequent phenomena).

[8] A somewhat larger corpus covering a very wide time span is the OED quotation database, estimated at 35M words [24]. Besides its modest size (and very few attestations of *to death* [33]), the OED quotations database is affected by balancing issues similar to those described for the Google Books Corpus.

[9] In the 1800 bin, no tokens are included from newspaper texts, and 82 out of 800 tokens (10.25%) were found in magazine texts.

[10] Additionally, even diachronic trends in balanced diachronic corpora may in a strict sense also be artefacts, as genres and registers are also subject to change. With respect to newspaper and magazine text, for instance, it has been shown that the changing "readerships and purposes of magazines versus newspapers result in different historical-linguistic patterns of use" [3].

6.2. Minimizing manual interference

The discussion of to what extent manual interference is needed or desirable in case-driven studies of semantic change is far from trivial, and, ultimately still undecided. In the spirit of the 'data-drivenness' discourse in preceding work [e.g. 41, 16], the procedure presented here aimed to minimize manual filtering and annotation – but such manual interference has not been entirely absent. In collecting the collocate verbs of *to death*, for instance, a substantial number of cases involved structures where the collocate of interest is not the verb *be* but its accompanying adjective. Similarly, cases where the verb form in closest proximity to *to death* (e.g. *we could prevent Scipio from* ***pummelling*** *the* ***dreaded*** *wizard to death*, COHA 1840) were corrected manually.

Furthermore, there are various points where further 'manual meddling' could be considered. In many instances that were retained in the dataset, *to death* has neither resultative nor intensifying meaning.[11] Given the limited relevance and potential effects on the output of the statistical analyses of irrelevant cases, it may be worth flagging or even excluding them from the dataset, as done in Margerie [33] and Perek [40]. Yet, such actions do involve elaborate manual annotation, and potentially introduce annotator judgments into the procedure that may diminish the 'data-driven' character of the study.[12]

Finally, with respect to the cluster analysis, a fully hands-off approach also implies that we trust the word embedding space to reflect meaningful semantic parameters, and that the resulting clusters capture, at least roughly, relevant properties of the underlying process. In the present case, it is reassuring to see that the narrative that emerges from our data analysis appears to align with what earlier linguistic research has proposed. However, it is not guaranteed that the verb clusters that were fed into the statistical analysis correspond with the semantic verb classes proposed in earlier research (e.g. actions of physical harm vs. mental verbs), or even with any groupings that are meaningful to humans. Additionally, while the bootstrapping procedure described in Section 4.1 renders the procedure more robust, it also makes it more difficult to examine which verbs constitute what cluster at which points in time. Given this lack of full transparency, the (as of yet unanswered) question becomes how to progress towards a method that reliably and robustly supports exploratory data analysis in cases not yet analyzed [also see 49, 20], and to what extent limiting manual involvement to an absolute minimum is warranted in specific case-driven studies.

7. Conclusion and Future Outlook

Drawing on the vast (and growing) body of computational research on semantic change, this study examined how computational models can be employed to track the host class expansion of grammaticalizing constructions, such as *to death*. By adjusting our scope to one specific and

[11] For instance, positive verb collocates such as *love* are attested earlier than expected in examples such as *He swore he wou'd love me to death* (EEBO, 1700), where *to death* most likely functions as a time adverbial ('he swore he'd love me until death') and not as a resultative ('he swore me he'd love me resulting in death') or an intensifier ('he swore he'd love me a lot'). These structures could, of course, have contributed to *to death*'s acquisition of intensifying meaning (loving someone until death implies loving them a lot), and hence be relevant to include. In other cases, however, the relevance of the query hit in relation to the semantic development described appears to be much less clear (e.g. *the first who turns his back to death*; EEBO, 1800).

[12] A possible, yet costly solution here would be to rely on multiple annotators, preferably with expertise in the historical language variety at hand [20].

relatively complex case, the procedure we outlined caters to case-driven research, which operates at a level of specificity and granularity that is not abundantly common in computational approaches to semantic change. Besides outlining the procedure, we flagged its current limitations and issues, which will hopefully entice further case-driven computational humanities research that will help reflect on and ultimately tackle the challenges that remain.

Acknowledgments

The work for this study has been made possible by the Platform Digital Infrastructure (Social Sciences and Humanities) fund (PDI-SSH). We want to thank the anonymous reviewers for their valuable suggestions, as well as Folgert Karsdorp for his advise regarding monotonic effects.

References

[1] R. Bamler and S. Mandt. "Dynamic word embeddings". In: *Proceedings of the 34th international conference on machine learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of machine learning research. Pmlr, 2017, pp. 380–389. URL: http://proceedings.mlr.press/v70/bamler17a.html.

[2] A. Betti, M. Reynaert, T. Ossenkoppele, Y. Oortwijn, A. Salway, and J. Bloem. "Expert Concept-Modeling Ground Truth Construction for Word Embeddings Evaluation in Concept-Focused Domains". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020, pp. 6690–6702. DOI: 10.18653/v1/2020.coling-main.586.

[3] D. Biber and B. Gray. "Being Specific about Historical Change: The Influence of Sub-Register". In: *Journal of English linguistics* 41.2 (2013), pp. 104–134.

[4] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

[5] Y. Bizzoni, S. Degaetano-Ortlieb, P. Fankhauser, and E. Teich. "Linguistic Variation and Change in 250 Years of English Scientific Writing: A Data-Driven Approach". In: *Frontiers in Artificial Intelligence* 3 (2020), p. 73. DOI: 10.3389/frai.2020.00073.

[6] R. A. Blythe and W. Croft. "S-curves and the mechanisms of propagation in language change". In: *Language* 88.2 (2012), pp. 269–304. DOI: 10.1353/lan.2012.0027.

[7] P. C. Bürkner. "Advanced Bayesian Multilevel Modeling with the R Package Brms". In: *R Journal* (2018). DOI: 10.32614/rj-2018-017.

[8] P. C. Bürkner and E. Charpentier. *Modeling Monotonic Effects of Ordinal Predictors in Bayesian Regression Models*. 2018. DOI: 10.31234/osf.io/9qkhj. URL: psyarxiv.com/9qkhj.

[9] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. "Stan: A probabilistic programming language". In: *Journal of statistical software* 76.1 (2017), pp. 1–32.

[10] D. Correia Saavedra. "Measurements of Grammaticalization: Developing a quantitative index for the study of grammatical change". PhD dissertation. Neuchâtel & Antwerpen: l'Université de Neuchâtel & Universiteit Antwerpen, 2019.

[11] M. Del Tredici, R. Fernández, and G. Boleda. “Short-term meaning shift: A distributional exploration”. In: *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 2069–2075. DOI: 10.18653/v1/N19-1210.

[12] D. Denison. “Log (ist) ic and simplistic S-curves”. In: *Motives for language change* 54 (2003), p. 70.

[13] H. Dubossarsky, S. Hengchen, N. Tahmasebi, and D. Schlechtweg. “Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 457–470. DOI: 10.18653/v1/P19-1044.

[14] H. Dubossarsky, D. Weinshall, and E. Grossman. “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*. Copenhagen, Denmark: Association for Computational Linguistics, 2017, pp. 1136–1145. DOI: 10.18653/v1/D17-1118.

[15] M. Faruqui, J. Dodge, S. K. Jauhar, C. Dyer, E. Hovy, and N. A. Smith. “Retrofitting Word Vectors to Semantic Lexicons”. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics, 2015, pp. 1606–1615. DOI: 10.3115/v1/N15-1184.

[16] M. Giulianelli, M. Del Tredici, and R. Fernández. “Analysing lexical semantic change with contextualised word representations”. In: *Proceedings of the 58th annual meeting of the association for computational linguistics*. Online: Association for Computational Linguistics, 2020, pp. 3960–3973. DOI: 10.18653/v1/2020.acl-main.365.

[17] E. Grave. *Language Identification*. 2017. URL: https://fasttext.cc/blog/2017/10/02/blog-post.html.

[18] W. L. Hamilton, J. Leskovec, and D. Jurafsky. “Cultural shift or linguistic drift? Comparing two computational measures of semantic change”. In: *Proceedings of the 2016 conference on empirical methods in natural language processing*. Austin, Texas: Association for Computational Linguistics, 2016, pp. 2116–2121. DOI: 10.18653/v1/D16-1229.

[19] W. L. Hamilton, J. Leskovec, and D. Jurafsky. “Diachronic word embeddings reveal statistical laws of semantic change”. In: *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)*. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 1489–1501. DOI: 10.18653/v1/P16-1141.

[20] S. Hengchen, N. Tahmasebi, D. Schlechtweg, and H. Dubossarsky. “Challenges for computational lexical semantic change”. In: Zenodo, 2021. DOI: 10.5281/zenodo.5040322.

[21] M. Hilpert and D. Correia Saavedra. “Using token-based semantic vector spaces for corpus-linguistic analyses: From practical applications to tests of theoretical claims”. In: *Corpus Linguistics and Linguistic Theory* 0.0 (2017). DOI: 10.1515/cllt-2017-0009.

[22] N. P. Himmelmann. *Lexicalization and grammaticization: opposite or orthogonal?" In What Makes Grammaticalization: A Look from Its Components and Its Fringes*. Ed. by W. Bisang, N. P. Himmelmann, and B. Wiemer. Berlin: Mouton de Gruyter, 2004.

[23] J. Hoeksema and D. Jo Napoli. "Just for the hell of it: A comparison of two taboo-term constructions". In: *Journal of Linguistics* 44.2 (2008), pp. 347–378. DOI: 10.1017/s002222670800515x.

[24] S. Hoffmann. "Using the OED quotations database as a corpus – a linguistic appraisal". In: *ICAME journal* 28 (2004), pp. 17–30.

[25] P. Hopper. "On some principles of grammaticalisation." In: *Approaches to grammaticalization*. Ed. by E. C. Traugott and B. Heine. Vol. 1. Amsterdam: John Benjamins, 1991, pp. 17–35.

[26] R. Hu, S. Li, and S. Liang. "Diachronic sense modeling with deep contextualized word embeddings: An ecological view". In: *Proceedings of the 57th annual meeting of the association for computational linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 3899–3908. DOI: 10.18653/v1/P19-1379.

[27] S. Jastrzebski, D. Leśniak, and W. M. Czarnecki. "How to evaluate word embeddings? on importance of data efficiency and simple supervised tasks". In: *arXiv preprint arXiv:1702.02170* (2017).

[28] A. Kroch. *Penn Parsed Corpora of Historical English*. Philadelphia, 2020. URL: https://www.ling.upenn.edu/hist-corpora/.

[29] A. Kutuzov, L. Øvrelid, T. Szymanski, and E. Velldal. "Diachronic word embeddings and semantic shifts: a survey". In: *Proceedings of the 27th international conference on computational linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018, pp. 1384–1397. URL: https://www.aclweb.org/anthology/C18-1117.

[30] G. Lorenz. "Really worthwhile or not really significant ?: A corpus-based approach to the delexicalization and grammaticalization of intensifiers in Modern English". In: *Typological Studies in Language*. Ed. by I. Wischer and G. Diewald. Vol. 49. Amsterdam: John Benjamins Publishing Company, 2002, pp. 143–161. DOI: 10.1075/tsl.49.11lor.

[31] Y. Luo, D. Jurafsky, and B. Levin. "From Insanely Jealous to Insanely Delicious: Computational Models for the Semantic Bleaching of English Intensifiers". In: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 1–13. DOI: 10.18653/v1/W19-4701.

[32] E. Manjavacas, Á. Kádár, and M. Kestemont. "Improving Lemmatization of Non-Standard Languages with Joint Learning". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 1493–1503. URL: https://www.aclweb.org/anthology/N19-1153.

[33] H. Margerie. "Grammaticalising constructions: to death as a peripheral degree modifier". In: *Folia Linguistica Historica* 45.Historica vol. 32 (2011). DOI: 10.1515/flih.2011.005.

[34] T. Mikolov, K. Chen, G. Corrado, and J. Dean. “Efficient Estimation of Word Representations in Vector Space”. In: *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. Ed. by Y. Bengio and Y. LeCun. 2013. URL: http://arxiv.org/abs/1301.3781.

[35] S. Mitra, R. Mitra, M. Riedl, C. Biemann, A. Mukherjee, and P. Goyal. “That’s sick dude!: Automatic identification of word sense change across different timescales”. In: *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)*. Baltimore, Maryland: Association for Computational Linguistics, 2014, pp. 1020–1029. DOI: 10.3115/v1/P14-1096.

[36] T. Nevalainen, H. Raumolin-Brunberg, J. Keränen, M. Nevala, A. Nurmi, M. Palander-Collin, A. Taylor, S. Pintzuk, A. Warner, et al. “Parsed Corpus of Early English Correspondence (PCEEC)”. In: *Oxford Text Archive Core Collection* (2006).

[37] E. A. Pechenick, C. M. Danforth, and P. S. Dodds. “Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution”. In: *Plos One* (2015), p. 24.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. “Scikit-learn: Machine learning in Python”. In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[39] J. Pennington, R. Socher, and C. Manning. “GloVe: Global Vectors for Word Representation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162.

[40] F. Perek. “Recent change in the productivity and schematicity of the *way* -construction: A distributional semantic analysis”. In: *Corpus Linguistics and Linguistic Theory* 14.1 (2018), pp. 65–97. DOI: 10.1515/cllt-2016-0014.

[41] F. Perek. “Using distributional semantics to study syntactic productivity in diachrony: A case study”. In: *Linguistics* 54.1 (2016). DOI: 10.1515/ling-2015-0043.

[42] R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik. *A Comprehensive Grammar of the English Language*. London: Longman, 1985.

[43] R. Rehurek and P. Sojka. “Software Framework for Topic Modelling with Large Corpora”. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (2010), pp. 45–50.

[44] A. Rosenfeld and K. Erk. “Deep neural models of semantic shift”. In: *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 474–484. DOI: 10.18653/v1/N18-1044.

[45] P. J. Rousseeuw. “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of computational and applied mathematics* 20 (1987), pp. 53–65.

[46] M. Rudolph and D. Blei. “Dynamic embeddings for language evolution”. In: *Proceedings of the 2018 world wide web conference*. Www ’18. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018, pp. 1003–1011. DOI: 10.1145/3178876.3185999. URL: https://doi.org/10.1145/3178876.3185999.

[47] E. Sagi, S. Kaufmann, and B. Clark. “Tracing semantic change with Latent Semantic Analysis”. In: *Current Methods in Historical Semantics*. Ed. by K. Allan and J. A. Robinson. Berlin, Boston: De Gruyter, 2011. DOI: 10.1515/9783110252903.161.

[48] D. Schlechtweg, B. McGillivray, S. Hengchen, H. Dubossarsky, and N. Tahmasebi. “SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection”. In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics, 2020, pp. 1–23. URL: https://aclanthology.org/2020.semeval-1.1.

[49] P. Sommerauer and A. Fokkens. “Conceptual Change and Distributional Semantic Models: an Exploratory Study on Pitfalls and Possibilities”. In: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 223–233. DOI: 10.18653/v1/W19-4728.

[50] K. Sun, H. Liu, and W. Xiong. “The evolutionary pattern of language in scientific writings: A case study of Philosophical Transactions of Royal Society (1665–1869)”. In: *Scientometrics* 126.2 (2021), pp. 1695–1724. DOI: 10.1007/s11192-020-03816-8.

[51] N. Tahmasebi, L. Borin, and A. Jatowt. “Survey of Computational Approaches to Lexical Semantic Change”. In: *arXiv:1811.06278 [cs]* (2019). URL: http://arxiv.org/abs/1811.06278.

[52] N. Younes and U.-D. Reips. “Guideline for improving the reliability of Google Ngram studies: Evidence from religious terms”. In: *Plos One* (2019), p. 17.

[53] L.-C. Yu, J. Wang, K. R. Lai, and X. Zhang. “Refining Word Embeddings for Sentiment Analysis”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, 2017, pp. 534–539. DOI: 10.18653/v1/D17-1056.