Grammar and meaning: Analysing the topology of diachronic word embeddings

Yuri Bizzoni
Saarland University
yuri.bizzoni@uni-saarland.de

Stefania Degaetano-Ortlieb
Saarland University
s.degaetano@mx.uni-saarland.de

Katrin Menzel
Saarland University
k.menzel@mx.uni-saarland.de

Pauline Krielke
Saarland University
mariepauline.krielke@uni-saarland.de

Elke Teich
Saarland University
e.teich@mx.uni-saarland.de

Abstract

The paper showcases the application of word embeddings to change in language use in the domain of science, focusing on the Late Modern English period (17-19th century). Historically, this is the period in which many registers of English developed, including the language of science. Our overarching interest is the linguistic development of scientific writing to a distinctive (group of) register(s). A register is marked not only by the choice of lexical words (discourse domain) but crucially by grammatical choices which indicate style. The focus of the paper is on the latter, tracing words with primarily grammatical functions (function words and some selected, poly-functional word forms) diachronically. To this end, we combine diachronic word embeddings with appropriate visualization and exploratory techniques such as clustering and relative entropy for meaningful aggregation of data and diachronic comparison.

1 Introduction

Word embeddings are by now a well established instrument for exploring and comparing corpora in terms of lexical fields and semantic richness (Lenci, 2008). More recently, diachronic word embeddings have been successfully applied to investigate lexical semantic change (e.g. Jatowt and Duh (2014); Hamilton et al. (2016a); Hellrich and Hahn (2016); Fankhauser and Kupietz (2017); Hellrich et al. (2018)). We supplement this line of work using diachronic word embeddings for the analysis of *change in grammatical use*, potentially indicating shifts in style/register. Word embeddings reflect shared usage contexts not only of lexical words but also of grammatical words. By grammatical words we understand function words (determiners, conjunctions, etc.) as well as some other specific word forms, such as *wh*-pronouns or *ing*-forms of verbs. Typically, the latter are poly-functional (e.g. verbal *ing*-forms can be gerunds, participles or markers of present continuous). Function words are high-frequency words and affected by change only in the long term (e.g. by becoming clitics or bound forms), while lexical words, typically in the lower frequency band, tend to change (meaning) fast. If pressure arises for grammar to change (e.g. for more economical expression), it will likely affect the poly-functional word forms first, which can spread to new syntagmatic environments or attract new lexemes and extend paradigmatically (like lexical words, unlike function words). To capture such developments, we employ diachronic word embeddings with visualization of word clusters on a diachronic axis combined with some other exploratory techniques, such as clustering and relative entropy. For instance, spread of a word/word form will result in the word moving in the overall embedding space, or paradigmatic extension will result in locally higher populate, denser spaces. Comparing lexical words, function words and poly-functional word forms, we inspect the overall topology of the embedding space over time as well as capture the internal composition of (selected) individual sub-spaces.

As a data set we use the Royal Society Corpus (RSC) (Kermes et al., 2016), a diachronic corpus of the Philosophical Transactions and the Proceedings of the Royal Society of London, which includes text material that is linguistically well explored in terms of style, register and diachrony (e.g. Biber and Finegan (1997); Atkinson (1999); Banks (2008); Degaetano-Ortlieb et al. (2018)).

Following related work (Section 2), we present our data and methods (Section 3). In Section 4, we analyze the embedding space in terms of change

Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pages 175–185
Florence, Italy, August 2, 2019.

in overall topology as well as changes in selected clusters. Zooming in on *ing*-forms, we also micro-inspect their (changing) syntagmatic contexts. We conclude with a summary and future work directions (Section 5).

2 Related work

Quantitative corpus-based approaches to language change (e.g. Hilpert (2006); Geeraerts et al. (2011); Sagi et al. (2011); Hilpert and Gries (2016)) share the basic assumption that language use is governed by statistical properties of lexical and grammatical items. In recent years, distributional semantic approaches based on word embeddings, often combined with clustering, capture this assumption in a bottom-up fashion, allowing to model semantic similarity of words from corpora. Approaches such as `word2vec` (Mikolov et al., 2013) and `SVD_PPMI` (Levy and Goldberg, 2014; Levy et al., 2015) trained on corpora covering several time spans allow investigating changes in the semantic usage of lexical items over time (Jatowt and Duh, 2014; Kim et al., 2014; Kulkarni et al., 2015; Hamilton et al., 2016a; Hellrich and Hahn, 2016). To also capture syntactic information, approaches have been developed that account for word order based on structured skip-gram models (Ling et al., 2015) and clustering the model output (Dubossarsky et al., 2015; Fankhauser and Kupietz, 2017).

Particularly targeted at the digital humanities as well as socio-historical corpus-linguistics are approaches which also allow meaningful ways to inspect the data. For instance, Hellrich et al. (2018) provide a visualization website (JeSemE) to inspect change in word meaning over time by means of line and bar plots considering different comparative parameters (word similarity, word emotion, typical context, and relative frequency); Fankhauser and Kupietz (2017) provide a visualization of change in the distributional semantics of words combined with their relative frequency over time.

3 Data and Methods

3.1 Data

As a data set we use v4.0 of the Royal Society Corpus (RSC)[1], containing the publications of the Philosophical Transactions and Proceedings of the Royal Society of London from 1665 to 1869 (ca. 32 million tokens and 10,000 documents). The RSC contains various types of metadata (e.g. author, publication date, text title) and linguistic annotations (e.g. lemma, parts of speech, sentence boundaries). Table 1 gives further statistics on the corpus.

decade	tokens	lemma	sentences
1660-69	455,259	369,718	10,860
1670-79	831,190	687,285	17,957
1680-89	573,018	466,795	13,230
1690-99	723,389	581,821	17,886
1700-09	780,721	615,770	23,338
1710-19	489,857	383,186	17,510
1720-29	538,145	427,016	12,499
1730-39	599,977	473,164	16,444
1740-49	1,006,093	804,523	26,673
1750-59	1,179,112	919,169	34,162
1760-69	972,672	734,938	27,506
1770-79	1,501,388	1,146,489	41,412
1780-89	1,354,124	1,052,006	37,082
1790-99	1,335,484	1,043,913	36,727
1800-09	1,615,564	1,298,978	45,666
1810-19	1,446,900	1,136,581	42,998
1820-29	1,408,473	1,064,613	43,701
1830-39	2,613,486	2,035,107	81,500
1840-49	2,028,140	1,565,654	70,745
1850-59	4,610,380	3,585,299	146,085
1860-69	5,889,353	4,474,432	202,488
total	31,952,725	24,866,457	966,469

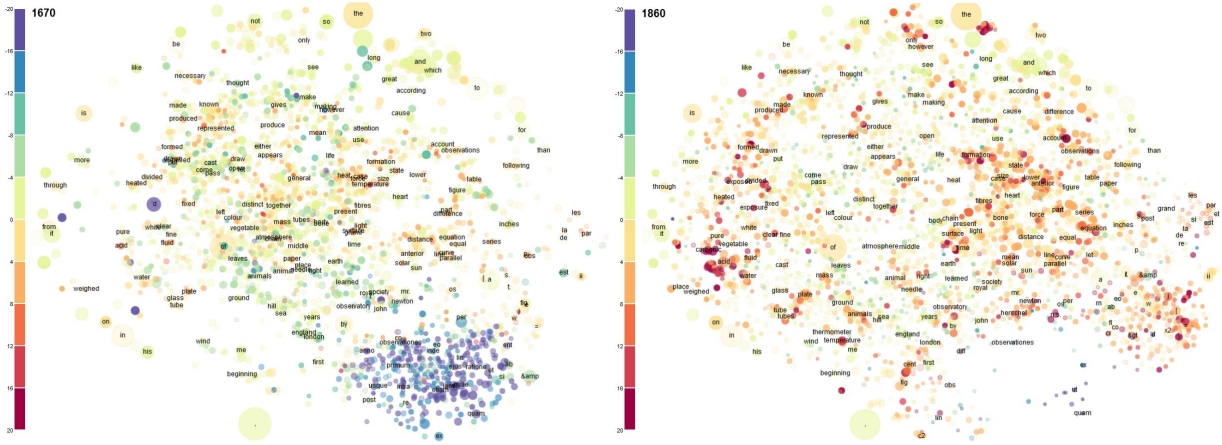
Table 1: Corpus statistics of the RSC per decade

3.2 Diachronic word embeddings

For computing word embeddings on a diachronic corpus, we follow the approach of Fankhauser and Kupietz (2017) – based on the structured skip-gram method described in Ling et al. (2015) with a one-hot encoding for words as input layer, a 200-dimensional hidden layer, and a window of [-5,5] as the output layer. Importantly, as this approach takes into account word order, it will capture grammatical patterns in word usage.

Word embeddings are calculated for each decade of the RSC. The embeddings for the first decade are initialized with a first-run training on the whole corpus, and subsequently refined for each decade of the 20 decades taken into consideration (1670–1860). The vocabulary of the models consists of a total of 117.165 100-dimensional points. The vocabulary consists only of "spaced" tokens (i.e. divided by space or punctuation in the original text). Multiword expressions and phrases are not taken into account, to maintain the original modelling as agnostic as possible about the content of the corpus. The models were

[1] Available open source at `http://fedora.clarin-d.uni-saarland.de/rsc_v4/`



(a) 1670-1679 (b) 1860-1869

Figure 1: Diachronic word embeddings of the 1670s and 1860s decades of RSC. Color denoting increasing (red) and decreasing (green) frequency. Size of the bubbles denoting relative frequency.

trained on non-lemmatized text. For interpretability, Fankhauser and Kupietz apply dimensionality reduction using t-Distributed Stochastic Neighbor Embedding by Maaten and Hinton (2008). Finally, a dynamic, interactive visualization of the resulting embeddings is provided which covers two crucial factors involved in diachronic change: *frequency* (encoded by colour – shades of violet-blue for decreasing frequency, shades of red-orange for increasing frequency) and *similarity in context of use* (encoded by proximity in space). For an example see Figure 1a. This allows us to explore changes in word use as shown in Section 4. As in most studies regarding distributional semantics, we will use cosine distance to compute the similarity between words in the space.

3.3 Investigating change in grammatical use

The large majority of studies performed on diachronic corpora through embedding spaces focuses on lexical semantics: to analyze changes in the distance between specific words over time (Szymanski, 2017), to infer semantic changes between specific categories of words, e.g. words referring to specific objects or concepts (Recchia et al., 2016), or to model the development of new terms with respect to the existing "neighborhoods" to infer their emergent semantic profile (Gangal et al., 2017).

But embedding spaces can be used to go beyond the study of change in lexical meaning (Jenset, 2013; Perek, 2016; Lenci, 2011), as they capture, to varying degrees, both paradigmatic and syntagmatic properties of words[2]. The same methods used for lexical words can be applied to grammatical words (as defined in Section 1): measuring the distance of individual words from their neighbours, mapping the evolution from their original position in the space, the nearest neighbour similarity, the similarity to other specifically selected words etc. Operating on grammatical words in the same way in which we traditionally operate on lexical words can return interesting observations, exactly as happens studying lexical words. Polyfunctional grammatical words, such as *ing*-forms, are at the boundary between lexis and grammar and are therefore particularly interesting because they can give us insights on the interplay between lexis and grammar. We outline here two main phenomena pertaining to the interplay between lexical semantics and grammatical function in distributional spaces: (1) diachronic expansion of the space; (2) diachronic clustering of poly-functional words with *ing*-forms as an exemplary case.

Considering (1), we measure average distances of lexical and function words as well as poly-functional word forms. Average distance is the average of the mean distances of each word from the rest of the vocabulary. In addition, we consider the average distance between words within a group (henceforth: inner distance), and the aver-

[2]For example, verbs in the past tense have a tendency to cluster with other verbs in the past tense with similar semantic properties, and verbs in the present continuous have a tendency to cluster with other verbs in the present continuous. The "window" size and the type of distribution taken into consideration of course have an important role in magnifying or blurring this aspect of words' distributional profile.

age distance of the group from all other words in the space (henceforth: outer distance). Change in inner distance reflects how much the words remain close to each other or drift apart in meaning/usage. Change in outer distance reflects how much semantically similar words become more isolated from all other words, possibly indicating a trend towards more specialized meaning/usage.

Considering (2), we operate in two main steps: (i) We first explore the sub-space of *ing*-forms to see whether meaningful clusters of verbs can be suspected. We do this by simply looking at verbs that have near neighbours, setting a threshold for what we consider *near*[3]. Through this very simple system, we elaborate an idea of what *kinds* of verbs are likely to constitute the clusters we are interested in. (ii) Once we have formed a hypothesis about the structure of the sub-space, we run some fairly popular algorithms of clustering and compare their results with our predictions and interpretations. This double step rises from the conviction that unsupervised clustering algorithms require a hypothesis about the structure of the data to both set their parameters and interpret their results, and that such hypothesis has to be acquired through an exploration of the space.

3.4 Investigating syntagmatic context

For further insights, we inspect the syntagmatic context of selected clusters of *ing*-forms extracting part-of-speech ngrams preceding an *ing*. We then use relative entropy (here: pointwise Kullback-Leibler Divergence (KLD; Kullback and Leibler (1951); Fankhauser et al. (2014); Tomokiyo and Hurst (2003)) to measure how distinctive particular syntagmatic contexts are for particular time periods. This is performed for each inspected feature (in our case a syntagmatic context in terms of a part-of-speech ngram, e.g. preposition-noun-*ing*-verb) comparing two time periods, $T1$ and $T2$ (cf. Equation (1)).

$$D_{feature}(T1||T2) = p(feature|T1) log_2 \frac{p(feature|T1)}{p(feature|T2)} \quad (1)$$

Basically, the probability of a feature in a time period $T1$ ($p(feature|T1)$) is compared to that feature in time period $T2$ ($p(feature|T2)$), i.e. the ratio of $T1$ vs. $T2$. To obtain features distinctive of $T1$ the ratio is weighted with the probability of that feature in $T1$. To obtain features distinctive of $T2$, the ratio between $T2$ and $T1$ is calculated and weighted by the feature's probability in $T2$. Divergence is measured in bits of information: the higher the amount of bits, the more the feature is distinctive of a given time period.

4 Analyses

In the analysis, we inspect (1) changes in the overall topology of the embedding space over time, and (2) the development of *ing*-forms of verbs.

4.1 Topology of the overall embedding space over time

Figures 1a and 1b show the embedding spaces for the RSC's first (1670s) and last (1860s) full decades. Most function words (e.g. *the, and, from*) are isolated in both decades indicating their functional status. Lexical words (e.g. verbs, nouns, adjectives), instead, cluster in one large group in the middle. Considering diachronic development, apart from local clusters disappearing altogether (e.g. a cluster of Latin, marked in blue), a visible general trend is the expansion of the overall space to smaller, more spread out and more separated clusters. Thus, the distance between words seems to increase in general, possibly indicating a process of specialization at word level. We test this for three cases: all words, function words and two poly-functional word forms (*ing-* and *-ed* forms of verbs).

All words. Analysing the spaces diachronically, we find that most lexical words[4] tend to drift further from each other over time. This does not mean that they do not form lexico-semantic clusters, but the average distance of each word from both its nearest neighbours (inner distance) and every other word in the space increases (outer distance) (see again Figure 2). Considering different sets of words in the spaces' vocabulary, we observe the same phenomenon: the average distributional distance tends to increase, both within the group (inner distance), and between the group and the rest of the lexicon (outer distance). In Figure 2 we show how this trend is clearly detectable in our spaces, independently of the words' frequencies. It can also be noted that the low frequency

[3] We will use a dynamic threshold for this task (see Section 4.2.2)

[4] Here, lexical words are all words that are not conjunctions, prepositions or adpositions.

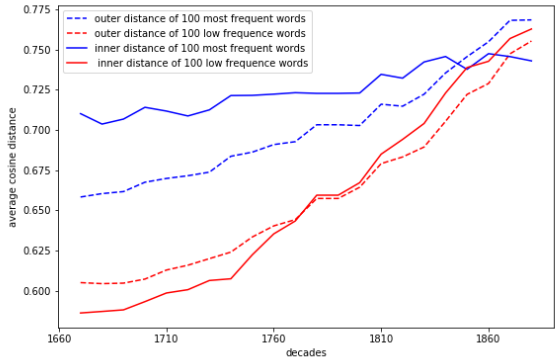


Figure 2: Diachronic increase of average distance of words. Average distance: mean of the mean distances of each word from the rest of the vocabulary. Selection of two groups: 100 high and 100 low frequency words in every decade. Inner distance: average distance between words within the group. Outer distance: average distance of the group from all other words in the space.

words maintain most of the time a lower average distance than high frequency words. We consider this a hint that the reason of the expansion of the space is due to *specialization*: the tail of the frequency curve tends to contain many highly technical words, with particularly specialized meanings. These words usually, while being far from the rest of the vocabulary, have a low number of very close neighbours, which represent those few words that happen to share similar specialized contexts. This is often considered an indication of single and specialized meaning (Hamilton et al., 2016b). In fact, words having a frequency lower than three in each decade have, on average, one neighbour which is considerably closer than the closest neighbour of highly frequent words (0.84 vs 0.71 cosine similarity on average). This all leads to the conclusion that the underlying mechanism is lexical specialization.

Function words. If we compare these general distributional behaviours to the behaviour of only function words (here: determiners, conjunctions and adpositions), we observe an interesting difference: function words tend to have an increasingly "reclusive" tendency. While their outer distance increases (see Figure 3), the inner distance stays stable. In other terms, while the average lexical word in our corpus undergoes a process of contextual specialization, function words do not.

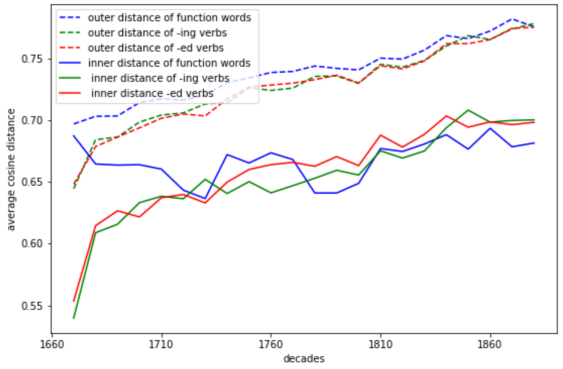


Figure 3: Inner and outer distances for function words, *ed*-verbs, and *ing*-verbs.

Poly-functional word forms. If lexical words undergo expansion in both directions (inner and outer distance), while function words only show an increase in the outer distance, we can assume that the increase in distances is due to the lexico-semantic side of words rather than their functional-grammatical side. This becomes particularly clear when we look at poly-functional word forms which share a common formal feature (e.g., suffix *ed*), but not a common semantic belonging. For example, the average inner distance between *ed*-forms of verbs[5], while increasing over time (see Figure 3), remains lower than their average outer distance: their grammatical side shows its effect on their distributional behaviour, somehow in tension with their semantic change. Among *ing*-forms of verbs, the same tension can be observed: the inner and outer distances both increase, but their inner distance remains smaller. Compare also trends in Figure 3, where the difference between inner and outer distance is immediately evident (outer distance always higher), with those in Figure 2, where such difference does not seem to retain a particular importance. See also Figure 4 for an exemplification of this semantic–grammatical tension.

4.2 Tracing the development of *ing*-forms

We have observed that for poly-functional word forms, which are very much "in between" lexis and grammar, inner distance grows more slowly. To analyze this phenomenon in more detail, we focus on *ing*-forms of verbs.

[5]We operate here under the somewhat simplistic assumption that verbs ending in *ed* represent the majority of past tenses.

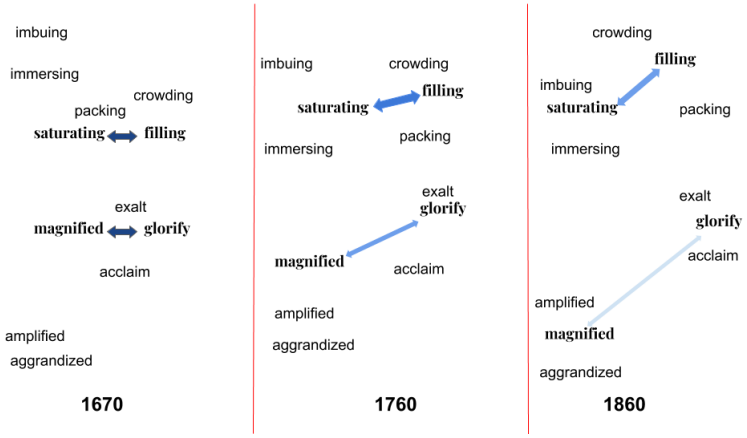


Figure 4: Example of semantic—grammatical tension. Two couples of verbs undergoing a semantic diversification (the left-side verbs become more specialized in meaning). In the lower side of the space, the two verbs have both semantic and grammatical differences. In the upper side of the space, the verbs have a growing semantic distance, but their grammatical profile remains similar; thus their distance grows more slowly.

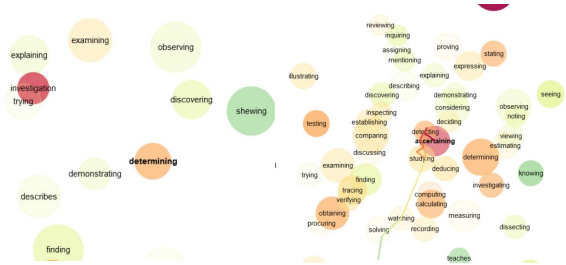
4.2.1 Diachronic frequency distribution of *ing*-forms

In a first step, to obtain a better understanding of the frequency distribution of *ing*-verb forms in the RSC corpus, we extract all verbs part-of-speech tagged as "gerunds or present participles" (VVG, VBG, VHG). Verbs with this tag include progressives, but exclude other verbs ending in - *ing* (e.g. *sing*, *bring*) or other parts of speech (e.g. *morning*, *spring*). We observe a fairly stable diachronic tendency. In addition, scientific writing is known to use *ing*-verbs most prominently as gerunds and participles rather than progressives (Biber et al., 1999). Indeed, the progressive form (i.e. BE + *ing*-verb) is quite infrequent in the RSC overall and it is declining over time; i.e. 250 occurrences of progressive per million tokens in the 1860s in 13,000 occurrences/million of ing-*forms* altogether.

4.2.2 Inspecting clusters of *ing*-forms

We consider all *ing*-forms per decade and consider as a cluster all neighbours closer than a given threshold distance. In this way, we can analyze (1) how close to other words *ing*-forms are on average, (2) how large their average cluster is (i.e. no. of words in a cluster), and (3) how much they tend to cluster with each other (i.e. whether and which *ing*-forms tend to occur in other *ing*-forms' neighbourhoods).

To build clusters we use a *dynamic* threshold. We set this threshold empirically to the decade's average distance of the nearest neighbours + .05. Thus, for each decade we can see which *ing*-forms have the highest number of "near" neighbours, and how many large clusters are formed, despite the general expansion of the space. From this exploratory analysis we observe that, first, despite our dynamic threshold, the density (i.e. number of words per cluster) of *ing*-clusters diminishes over time. We ascribe this effect, like the more general expansion of the space, mostly to the lexical-semantic component of the verbs involved: their meaning becomes more specific, their context more specialized – and thus less overlap between their contexts is observed. At the same time, the words that are at the *center* of a cluster (i.e. words with relatively large and close neighbourhoods) appear to belong to three increasingly distinct categories.



(a) 1670s (b) 1860s

Figure 5: Academic *ing*-verbs in the RSC

The most prominent category are so-called *academic verbs*, such as *ascertaining*, *determining*, *examining* etc. acquiring relatively tight and large neighbourhoods (see Figure 5a and 5b[6]).

[6]Figure 5b showing the diachronic trajectory of *ascertaining* moving towards the center of the cluster. Color of the trajectory denoting frequency (green: lower/red: higher)

The complementary analysis of the most frequent neighbours (words that occur most frequently in other words' close neighbourhood) shows the same phenomenon: academic verbs rise in frequency. The two other main categories we observe at the center of large clusters are *change-of-state* verbs (*saturating, diluting*, etc.) and *motion* verbs (*passing, falling*, etc.).

4.2.3 Clustering specialized vs. broader meanings

Based on the above findings, which gave us a general idea of possible clusters, we can now apply some traditional clustering algorithms to our dataset. We will show the results of three algorithms: Affinity Propagation (AP) (Frey and Dueck, 2007), DBSCAN (Ester et al., 1996; Tran et al., 2013), and MiniBatch K Means (Sculley, 2010; Feizollah et al., 2014). Results are presented in Table 2.

Affinity Propagation, much like DBSCAN, does not require a pre-determined number of clusters, i.e. it defines its own number of centroids. While usually seen as an advantage, in our case it could result in a flaw: these algorithms tend towards a micro-clustering (clustering tight relationships), leading to many small clusters of specialized meanings of *ing*-verbs. This would probably shadow the larger and looser clustering resulting from a possible interplay of semantics with more grammatical classes of *ing*-verbs. In fact, Affinity Propagation individuates a large number of *ing*-clusters, and most relevant, an *increasing* number of *ing*-clusters over time. What we see here is lexico-semantic specialization at work: every cluster contains "few" words semantically very close, e.g. *drawing - tracing, preceding - foregoing*.

DBSCAN does not require a pre-determined number of clusters either, but a fixed threshold and a fix minimum of neighbours to consider members of a cluster. While the number of centroids is lower than the number found by Affinity Propagation, it still increases over time.

Unlike the previous two algorithms, MiniBatch K Means requires a heavier pre-interpretation of the data: we need to know how many clusters we are looking for. While usually seen as a disadvantage, once we have more than an educated guess – thanks to our previous exploration of the data – it can turn into a strength: we can force the algorithm to look beyond the most evident micro-clusters and define a larger subdivision of the space. In fact, once we use the K Means algorithm on the *ing*-subspace, setting the number of centroids to 3 (the number of verb classes we have observed through our exploration in Section 4.2.2), we obtain results that are very close to our observations. The verbs falling in the three groups more and more pertain to what we would call academic, change-of-state, and motion verbs (see Table 2, 1860s decade). The centroids determined by the MiniBatch K Means algorithm for these three clusters grow further apart through time, and especially from the beginning of the 19th century we can detect a growing distributional difference between the three centroids of these clusters.

4.2.4 Grammatical classes of *ing*-clusters

To observe whether the use of these main *ing*-clusters differs in terms of grammatical class (gerund vs. participle), we further inspect their syntagmatic context. For this, we generate lists of the top 30 verbs derived from the clusters and extract their preceding part-of-speech ngrams to observe how their use varies in syntactic context. Using Kullback-Leibler Divergence we can inspect which possible grammatical classes (i.e. gerund vs. participle) are distinctive of later time periods in comparison to earlier time periods considering each semantic group of verbs (i.e. academic, change-of-state, motion).

Figure 6 shows the frequency distribution of the three clusters across decades in the RSC. Change-of-state verbs (e.g. *purifying, warming, cooling*) seem to remain relatively stable, showing only a very slight increase. Motion verbs (e.g. *passing, extending, running*) increase especially after 1820. Verbs belonging to the academic semantic sub-space rise until 1810 and decline afterwards. It seems that the beginning of the 19th century (1810-1840) marks a period of change.

Using relative entropy, we compare the part-of-speech ngrams of the three main clusters (academic, motion, and change-of-state verbs in *ing*-form) for the period preceding the 1810s and the period after the 1840s (i.e. 1660-1810 vs. 1850-1869). Table 3 shows the top five ngrams for each cluster, ranked by KLD. By inspecting the grammatical class of each ngram, we see a clear difference between the academic and the motion clusters: while verbs in the academic *ing*-cluster are used as gerunds, those in the motion *ing*-cluster are used as participles. Change-of-state *ing*-verbs are also most distinctively used as gerunds. This

Decade	Affinity Propagation (AP)	DBSCAN	Minibatch KMeans
1660	*Extending, reaching, proceeding.* *Crying, coughing, sweating.* *Shading, scattering, tracing.*	*Abounding, according, adding.* *Whiting, widening, willing.*	*Detaching, wetting, squeezing.* *Verifying, deciding, transferring.* *Playing, retiring, accumulating.*
1760	*Pricking, stimulating, snapping.* *Following, lowing, preceding.* *Informing, troubling, acquainting.*	*Abating, abounding, abstracting.* *Lessening.* *Deducting, subtracting, weighing.*	*Arranging, attaching, immersing.* *Arranging, studying, illustrating.* *Interlacing, arranging, transforming.*
1860	*Nourishing, binding, imbibing.* *Snapping, widening, pricking.* *Stimulating, promoting, biting.*	*Abounding, absorbing, abstracting.* *Integrating, introducing, putting.* *Arching, running, sweeping.*	*Determining, establishing, studying.* *Passing, extending, running.* *Purifying, agitating, warming.*

Table 2: Clusters of *ing*- forms with AP, DBSCAN and KMeans.

POS ngram	class	relative entropy (KLD)	example
Academic verbs			
SENT.IN.VVG	Gerund	0.0620	***. In examining*** *the laws*
VVN.IN.VVG	Gerund	0.0587	*the formulae* ***employed in finding*** *these logarithms*
NN.IN.VVG	Gerund	0.0492	*Potasse for the* ***purpose of ascertaining*** *whether*
IN.RB.VVG	Gerund	0.0183	*opportunity* ***of sufficiently investigating*** *the errors*
SENT.RB.VVG	Gerund	0.0110	***. Hence considering*** *an equation*
Motion verbs			
JJ.NN.VVG	Participle	0.0412	*the* ***smaller extremity lying*** *in contact with*
(.,.VVG	Participle	0.0370	*the tangential force (F)****, forming*** *two equal*
JJ.NNS.VVG	Participle	0.0362	*refracting the* ***visual rays passing*** *thorough them*
IN.NNS.VVG	Participle	0.0327	*dark cloud* ***of ashes falling*** *from the volcano*
SENT.IN.VVG	Gerund	0.0270	***. After passing*** *the central layer*
Change-of-state verbs			
VVN.IN.VVG	Gerund	0.1116	*more strongly* ***magnetized by placing*** *them*
SENT.IN.VVG	Gerund	0.0630	***. By heating*** *it to above the boiling*
VVZ.IN.VVG	Gerund	0.0590	***crystallizes on cooling***
NN.,.VVG	Participle	0.0254	*a deep oblique* ***fold , penetrating*** *from the inner side*
JJ.NN.VVG	Participle	0.0235	*the* ***chylo-aqueous fluid filling*** *the ciliated*

IN: preposition, JJ: adjective, NN(S): common noun (pl.), RB: adverb, SENT: full stop, VVG: *ing*-form, VVN: participle, VVZ: present tense

Table 3: Top five part-of-speech ngrams of each verb cluster distinctive for the 1850s period (1850-69 vs. 1660-1800)

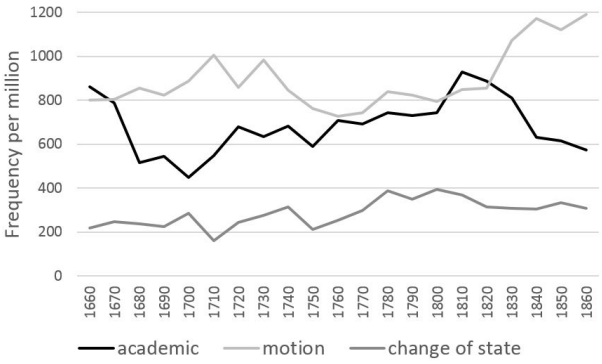


Figure 6: Frequency distribution of main clusters in the RSC

shows that besides capturing semantic relatedness, the diachronic word embeddings also capture grammatical use.

5 Conclusion

We have shown an analysis of diachronic word embeddings based on a diachronic corpus of English scientific writing. The aim of the analysis has been to trace changes in the embeddings of words with grammatical functions (function words, polyfunctional word forms) compared to lexical words. Analyzing the changing topology of the embedding space over time, operating with the notions of inner and outer distance (see Section 3), we were able to show that grammatical words behave differently from lexical words (Section 4). Specifically, we focused on words that have both a lexical meaning and specific grammatical functions, exemplified by *ing*- and *ed*-forms of verbs, because it seemed to us that such forms are common hosts for short to mid-term change in language use in scientific language. Here, we showed that *ing*-forms of verbs form three semantic groups (academic, motion and change-of-state), where change-of-state and academic verbs tend to be gerunds and motion verbs tend to be used as participles.

Methodologically, we showed that diachronic word embeddings are well suited to detect change not only in lexical but also in grammatical use as well as the interplay of lexis and grammar. Di-

achronic word embeddings combined with informative visualization and appropriate exploratory techniques (here: clustering and relative entropy) presents a powerful tool to investigate changing language use.

In our future work, we plan to inspect other poly-functional words and word forms, such as *wh*-words, because they seem to be involved in the development of scientific style as well. At the level of lexical words, we plan to analyze the embedding space in terms of domain-specific vocabulary. As mentioned in our analyses in various places, the overall trend in scientific vocabulary is specialization. To form distinctive registers (e.g., the language of chemistry, physics, medicine, etc.), vocabulary needs to become diversified. To track diversification related to register formation is therefore a high priority on our research agenda.

Acknowledgments

This work has been supported by *Deutsche Forschungsgemeinschaft* (DFG) under grant 'SFB 1102: Information Density and Linguistic Encoding'. We thank Peter Fankhauser for the initial implementation and visualization of word embeddings for the RSC.

References

Dwight Atkinson. 1999. *Scientific Discourse in Sociohistorical Context: The Philosophical Transactions of the Royal Society of London, 1675-1975*. Erlbaum, New York.

David Banks. 2008. *The Development of Scientific Writing: Linguistic Features and Historical Context*. Equinox, London/Oakville.

Douglas Biber and Edward Finegan. 1997. Diachronic Relations among Speech-based and Written Registers in English. In Terttu Nevalainen and Leena Kahlas-Tarkka, editors, *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*, pages 253–276. Société Néophilologique, Helsinki.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Longman, Harlow, UK.

Stefania Degaetano-Ortlieb, Hannah Kermes, Ashraf Khamis, and Elke Teich. 2018. An Information-Theoretic Approach to Modeling Diachronic Change in Scientific English. In Carla Suhr, Terttu Nevalainen, and Irma Taavitsainen, editors, *From Data to Evidence in English Language Research*, Language and Computers. Brill, Leiden.

Haim Dubossarsky, Yulia Tsvetkov, Chris Dyer, and Eitan Grossman. 2015. A Bottom Up Approach to Category Mapping and Meaning Change. In *Proceedings of the NetWordS Final Conference*, pages 66–70, Pisa, Italy.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*, volume 96, pages 226–231.

Peter Fankhauser, Jörg Knappen, and Elke Teich. 2014. Exploring and Visualizing Variation in Language Resources. In *Proceedings of the 9th LREC*, pages 4125–4128, Reykjavik, Iceland. ELRA.

Peter Fankhauser and Marc Kupietz. 2017. Visualizing Language Change in a Corpus of Contemporary German. In *Proceedings of the Corpus Linguistics International Conference*, Birmingham, UK.

Ali Feizollah, Nor Badrul Anuar, Rosli Salleh, and Fairuz Amalina. 2014. Comparative Study of K-Means and Mini Batch K-Means Clustering Algorithms in Android Malware Detection using Network Traffic Analysis. In *2014 International Symposium on Biometrics and Security Technologies (ISBAST)*, pages 193–197. IEEE.

Brendan J Frey and Delbert Dueck. 2007. Clustering by Passing Messages between Data Points. *Science*, 315(5814):972–976.

Varun Gangal, Harsh Jhamtani, Graham Neubig, Eduard Hovy, and Eric Nyberg. 2017. Charmanteau: Character Embedding Models for Portmanteau Creation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2907–2912, Copenhagen, Denmark. ACL.

Dirk Geeraerts, Caroline Gevaert, and Dirk Speelman. 2011. How Anger Rose: Hypothesis Testing in Diachronic Semantics. *Current Methods in Historical Semantics*, 73:109.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural Shift or Linguistic Drift? Comparing Two Computational Models of Semantic Change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, Texas.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1489–1501, Berlin, Germany. ACL.

Johannes Hellrich, Sven Buechel, and Udo Hahn. 2018. JESEME: A Website for Exploring Diachronic Changes in Word Meaning and Emotion. *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations (COLING 2018)*, pages 10–14.

Johannes Hellrich and Udo Hahn. 2016. Measuring the Dynamics of Lexico-Semantic Change Since the German Romantic Period. In *Digital Humanities (DH)*, pages 545–547.

Martin Hilpert. 2006. Distinctive Collexeme Analysis and Diachrony. *Corpus Linguistics and Linguistic Theory*, 2(2):243–256.

Martin Hilpert and Stefan Th Gries. 2016. Quantitative Approaches to Diachronic Corpus Linguistics. *The Cambridge Handbook of English Historical Linguistics*, pages 36–53.

Adam Jatowt and Kevin Duh. 2014. A Framework for Analyzing Semantic Change of Words across Time. In *Proceedings of Digital Libraries Conference (JCDL 2014 / TPDL 2014)*, pages 229–238, London, UK. ACM Press.

Gard B Jenset. 2013. Mapping Meaning with Distributional Methods: A Diachronic Corpus-based Study of Existential there. *Journal of Historical Linguistics*, 3(2):272–306.

Hannah Kermes, Stefania Degaetano-Ortlieb, Ashraf Khamis, Jörg Knappen, and Elke Teich. 2016. The Royal Society Corpus: From Uncharted Data to Corpus. In *Proceedings of the 10th LREC*, Portorož, Slovenia. ELRA.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and 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*, pages 61–5, Baltimore, Maryland, USA.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically Significant Detection of Linguistic Change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635, Florence, Italy.

Solomon Kullback and Richard A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Alessandro Lenci. 2008. Distributional Semantics in Linguistic and Cognitive Research. *Italian Journal of Linguistics*, 20(1):1–31.

Alessandro Lenci. 2011. Composing and Updating Verb Argument Expectations: A Distributional Semantic Model. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 58–66, Portland, Oregon, USA. ACL.

Omer Levy and Yoav Goldberg. 2014. Neural Word Embedding as Implicit Matrix Factorization. In *Advances in Neural Information Processing Systems*, pages 2177–2185.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Wang Ling, Chris Dyer, Alan W Black, and Isabel Trancoso. 2015. Two/too Simple Adaptations of word2vec for Syntax Problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1299–1304, Denver, Colorado. ACL.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, USA. ACL.

Florent Perek. 2016. Using Distributional Semantics to Study Syntactic Productivity in Diachrony: A Case Study. *Linguistics*, 54(1):149–188.

Gabriel Recchia, Ewan Jones, Paul Nulty, John Regan, and Peter de Bolla. 2016. Tracing Shifting Conceptual Vocabularies through Time. In *European Knowledge Acquisition Workshop*, pages 19–28. Springer.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2011. Tracing Semantic Change with Latent Semantic Analysis. *Current Methods in Historical Semantics*, 73:161–183.

David Sculley. 2010. Web-scale K-Means Clustering. In *Proceedings of the 19th International Conference on World Wide Web*, pages 1177–1178, Raleigh, North Carolina, USA. ACM.

Terrence Szymanski. 2017. Temporal Word Analogies: Identifying Lexical Replacement with Diachronic Word Embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 448–453, Vancouver, Canada. ACL.

Takashi Tomokiyo and Matthew Hurst. 2003. A Language Model Approach to Keyphrase Extraction. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18*, MWE '03, pages 33–40, Stroudsburg, PA, USA. ACL.

Thanh N Tran, Klaudia Drab, and Michal Daszykowski. 2013. Revised DBSCAN Algorithm to Cluster Data with Dense Adjacent Clusters. *Chemometrics and Intelligent Laboratory Systems*, 120:92–96.