Word Embeddings: Reliability & Semantic Change

Johannes Hellrich

Ph.D. thesis
Friedrich Schiller University Jena

Available in print from IOS Press as
Volume 347 of Dissertations in Artificial Intelligence

ISSN 0941-5769 (print)
ISSN 2666-2175 (online)

https://www.iospress.nl/book/
word-embeddings-reliability-semantic-change

Zusammenfassung

Word Embeddings sind ein Verfahren der distributionellen Semantik und werden zunehmend zur Untersuchung von Wortwandel genutzt. Allerdings sind typische Erstellungsverfahren probabilistisch, was ihre Zuverlässigkeit und damit auch die Reproduzierbarkeit von Studien limitiert.
Ich habe dieses Problem sowohl theoretisch als auch experimentell untersucht und festgestellt, dass Varianten des SVD_{PPMI} Algorithmus davon unbetroffen sind.
Zusätzlich habe ich die JeSemE Webseite entwickelt, die diachrone Studien auf Basis von Word Embeddings ohne technisches Vorwissen ermöglicht. JeSemE bietet Zugriff auf Trends in Denotation und emotionaler Konnotation für fünf diachrone Korpora.
Meinen distributionellen Ansatz habe ich in zwei Fallstudien getestet, die sich mit der Geschichte der Elektrizitätsforschung und mit Wörtern von Bedeutung für die Epoche der Romantik auseinandersetzen. Sie haben gezeigt, dass distributionelle Methoden ein wertvolles Werkzeug für die digitalen Geisteswissenschaften sein können.

Abstract

Word embeddings are a form of distributional semantics increasingly popular for investigating lexical semantic change. However, typical training algorithms are probabilistic, limiting their reliability and the reproducibility of studies.
I investigated this problem both empirically and theoretically and found some variants of the SVD_{PPMI} algorithm to be unaffected.
Furthermore, I created the JeSemE website to make word embedding based diachronic research more accessible. It provides information on changes in word denotation and emotional connotation in five diachronic corpora.
Finally, I conducted two case studies on the applicability of these methods by investigating the historical understanding of electricity as well as words connected to Romanticism. They showed the high potential of distributional semantics for further applications in the digital humanities.

Acknowledgments

First of all, I want to thank Prof. Dr. Udo Hahn for his supervision and guidance during my time as a graduate student. He gave me the freedom to pursue my ideas while always providing support and motivating me to strive for high scientific standards and a new state-of-the-art.
Furthermore, I want to thank Prof. Dr. Holger Diessel for examining this thesis and providing me opportunities for discussion.
I am grateful to the Graduate School "The Romantic Model" (grant GRK 2041/1 from Deutsche Forschungsgemeinschaft) which provided funding and an interdisciplinary forum, as well as the EU for previous funding during the MANTRA project (EU STREP project grant 296410).
My thanks go to my current and former colleagues who were always ready for discussion and collaboration.
Special thanks for proofreading go to Sven Büchel, Luise Modersohn and Christoph Rzymski as well as Tinghui Duan, Bernd Kampe, Tobias Kolditz, Christina Lohr, Stephanie Luther, Alexander Stöger and my lovely Diana Elias.
Last but not least, I am deeply grateful to my parents Hermann & Ruth Hellrich—I could not have completed this thesis without your trust and support!

Contents

List of Tables

List of Figures

Important Abbreviations

CBOW Continuous Bag-of-Words, a word embedding algorithm, see Section 2.3.3.

COHA Corpus of Historical American English, see Section 3.1.

DTA Deutsches Textarchiv Kernkorpus ['German text archive core corpus'], see Section 3.2.

GBF English Fiction subset of the Google Books Ngram corpus, see Section 3.3.

GBG German sub-corpus of the Google Books Ngram corpus, see Section 3.3.

GloVe Global Vectors, a word embedding algorithm, see Section 2.3.4.

NN Artificial Neural Networks, a machine learning approach inspired by biological neurons, see Section 2.3.3.

PMI Pointwise Mutual Information, a word association measure, see Section 2.2.1.

PPMI Positive Pointwise Mutual Information (PMI), a variant of the PMI word association measure, see Section 2.2.1.

RSC Royal Society Corpus, see Section 3.4.

SG Skip-gram, a word embedding algorithm, see Section 2.3.3.

SGNS Skip-gram Negative Sampling, a more efficient variant of the SG word embedding algorithm, see Section 2.3.3.

SGHS Skip-gram Hierarchical Softmax, a more efficient variant of the SG word embedding algorithm, see Section 2.3.3.

SVD Singular Value Decomposition, a linear algebra method for reducing the dimensionality of data, see Section 2.3.2.

$\mathbf{SVD_{PPMI}}$ Singular Value Decomposition of a PPMI matrix, a word embedding algorithm, see Section 2.3.2.

$\mathbf{SVD_{wPPMI}}$ Singular Value Decomposition of a PPMI matrix with weighting-based downsampling, a word embedding algorithm, see Sections 2.3.2 & 4.6.

VAD Valence-Arousal-Dominance, a three-dimensional model for emotions, see Section 5.1.1.

$\boldsymbol{\chi^2}$ Pearson's χ^2, a word association measure, see Section 2.2.1.

Chapter 1

Introduction

Computational studies of lexical semantics and semantic change are increasingly popular (see e.g., Manning (2015), Kutuzov et al. (2018)), due to both the availability of large corpora containing up to 6% of all books ever published (Michel et al., 2011) and the development of **word embeddings**. Word embeddings (e.g., `word2vec` by Mikolov et al. (2013a,b)) represent lexical semantics with numerical vectors by observing the "company" each word "keeps" (Firth, 1968, p. 11), i.e., its co-occurrence patterns. They can be used to measure word similarity, transforming the long ongoing practice of corpus-based studies in (historical) linguistics (Biber et al., 2000; Hilpert & Gries, 2009; Kohnen, 2006).

However, as with every new method, researchers must make sure that results are both valid—they measure what they are intended to measure—and reliable—repeated measurements are consistent with each other (Carmines & Zeller, 1992, pp. 11–12). The latter, however, is lacking for most word embedding algorithms (Antoniak & Mimno, 2018; Chugh et al., 2018; Hellrich & Hahn, 2016b, 2017a; Pierrejean & Tanguy, 2018a; Wendlandt et al., 2018). These unreliable methods can mislead scholars, as data-driven analyses of large corpora—also known as 'distant reading' (Moretti, 2013)—are increasingly popular in the digital humanities and social sciences (e.g., Michel et al. (2011), Jockers (2013)). Unreliable methods might also affect business and governmental decisions by impeding, e.g., the automatic maintenance of knowledge resources (e.g., Klenner & Hahn (1994)) or the observation of trends in social media (e.g., Preoţiuc-Pietro et al. (2016) or Arendt & Volkova (2017)).

Automatic diachronic studies are not only impeded by unreliable methods, but also by barriers excluding many potential users. Up to now, the usage of word embeddings requires programming skills, non-trivial computational resources and sometimes also pay-to-use corpora. This complicates their use for many scholars as they lack one or more of these prerequisites. A potential solution are websites providing access to statistical analyses (e.g., Davies (2014), Jurish (2015)) and diachronic lexical semantics (Hellrich et al., 2018a; Hellrich & Hahn, 2017b) derived from word embeddings.
This thesis tackles the above-mentioned issues and provides two case studies on using word embeddings to investigate the history of science and words related to Romanticism. Thus, it is concerned with lexical semantics only and not with other diachronic questions such as the automatic delimitation of historical epochs (Popescu & Strapparava, 2013) or visualizing changes in lexical resources (Theron & Fontanillo, 2015).

1.1. Words and Meaning

The word *word*[1] is used ambiguously in this thesis to allow for fluent writing despite a mismatch between linguistic definitions[2] and technical possibilities. In general, it is used in the sense of 'token', i.e., an uninterrupted grapheme sequence as occurring in a corpus, as well as 'type', i.e., an element in the set of tokens for a corpus. For lemmatized corpora, all tokens can be assumed to be replaced with instances of a (pseudo-)lemma and all types can be assumed to be equivalent with lexemes. This thesis contains no word-sense-aware experiments, thus lexical units were never modeled. Words and not word clusters or concepts[3] were chosen to provide a fixed starting point for analyses.

[1] Throughout this thesis, object language will be given in *italics*.

[2] See for example the SIL Glossary: `https://glossary.sil.org/` [Accessed May 28th 2019].

[3] Concepts or, more accurately, models approximating these mental representations, can be used to organize observations and describe historical developments (Kuukkanen, 2008; Thagard, 1990). However, that line of research is not concerned with language itself or a link between language and other phenomena, but with changes in the way these phenomena are defined.

Word embeddings are a shallow neo-structuralist[4] (Geeraerts, 2010) model of word meaning, i.e., describing words by patterns in their usage according to a "meaning-is-use theory" (Lyons, 1996, p. 40). They achieve a high correlation with human judgments when tested for their ability to measure similarity (Levy et al., 2015). More complex formal approaches for modeling word meaning would assume a hierarchy, grounding in logic or some kind of semantic primitives (see e.g., Grefenstette (1994, ch. 2) or Jurafsky & Martin (2009, chs. 17–20)). Such approaches were avoided in this thesis, as they would lead to additional levels of abstraction and thus potential artifacts during data analysis.
From a lingusitic point of view, a word's meaning contains several aspects (see e.g., Lyons (1996)). In the following all subjective-emotional aspects will be referred to as connotation, while other aspects will be referred to as denotation, even if they are not strictly referential. For example, according to this definition *shrink* and *psychotherapist* have the same denotation, but the former has a different (i.e., negative) connotation. It is possible to quantify emotional connotation with several models (Bradley & Lang, 1994; Ekman, 1992). Distributional information can then be used to predict word emotions, even in a diachronic setting (Cook & Stevenson, 2010; Turney & Littman, 2003).

1.2. Semantic Change

How languages change is a core question of linguistics and affects all linguistic levels.[5] Probably the most popular research topic is not semantic change, but sound change which has been used since the late 18$^{\text{th}}$ century to study the genealogical relationships between languages (e.g., Collinge (1990), Hock (1991, ch. 20)).
Semantic change can take many forms (Blank, 1999; Bloomfield, 1984; Hock, 1991) which can be roughly grouped into the following three categories:[6]

[4] Structuralist only from a linguistic, but not from an artificial intelligence point of view. According to the latter, representations based on non-localized patterns (such as word embeddings) are connectionist (Hinton, 1986).

[5] Some changes even link different levels, e.g., a taboo against uttering a word may lead to voluntary mispronunciations which can become codified (Hock, 1991, pp. 296–297).

[6] Examples from Bloomfield (1984, pp. 426–427).

Widening & Narrowing describe words becoming more general or more specific in their denotation. An example for widening is *dog*, being derived from *dogge*, the name of a dog breed in Middle English. Another canine example, Old English *hund*, lost its general meaning of 'dog' and became *hound* 'hunting dog'.

Elevation & Degeneration describe words becoming more or less positive in their emotional connotation and even denotation (e.g., social status). For example, Old English *cniht* and *cnafa* both meant 'boy, servant'. The former was elevated and became *knight*, whereas the latter was degenerated to the modern *knave*.

Metaphoric use in a very broad sense—including metonymy, synecdoche, hyperbole, litotes and euphemism (Hock, 1991, p. 285) which many treat separately (e.g., Bloomfield (1984, ch. 24) or Blank (1999))—describes words being used in a non-literal yet codified manner. For example, Old English *cēace* 'jaw' became *cheek* and *bitter* is derived from Germanic **[bitraz]* 'biting'.

Semantic change can be caused by both intra-linguistic and extra-linguistic developments, making its study salient to linguists and other scholars alike. In the former case, speakers of different varieties interact and mix their usage (Schmidt, 2007, pp. 3–5) or learners acquire words with a too specific or too general meaning due to repeated miscommunication (Bloomfield, 1984, ch. 24). In the latter case, cultural or technological developments necessitate a change in meaning, e.g., words related to livestock being repurposed with the spread of currency (Bloomfield, 1984, pp. 435–436). This can be due to a change in the environment, e.g., a sense being added to *mouse* due to the development of a vaguely rodent-shaped input device, or due to a change of environment, e.g., Spanish varieties in Europe and America use *léon* for 'lion' respectively 'puma', depending on which large predatory cat is local (Blank, 1999).

1.3. Reliability

Reliability and validity are two main criteria that need to be fulfilled in empirical research. While the former describes how much an experiment is affected by random errors, the latter is concerned with non-random biases (Carmines & Zeller, 1992, pp. 11–15).

Unreliable methods lead to experiments which cannot be properly repeated. However, repeatable experiments are vital for scientific progress as they enable others to test claims and extend existing work (see e.g., Mesirov (2010), Ivie & Thain (2018), Open Science Collaboration (2015)). Most word embedding algorithms are probabilistic and thus inherently unable to produce the same results in repeated experiments, unless their random processes are made deterministic which can, however, distort experimental results (Henderson et al., 2018).
So far, the question of reliability for word embedding experiments was mostly ignored, probably due to embeddings being mostly used as features in larger systems involving further probabilistic processes. In contrast, corpus linguistic analyses are commonly done with perfectly reliable statistical methods—a high standard that should be preserved for the increasingly popular application of word embeddings as a novel form of corpus linguistics (see e.g., Jo (2016), Hamilton et al. (2016c), Kulkarni et al. (2016)).

1.4. Contributions

The main contribution of this thesis lies in the exploration of word embedding reliability. Reliability problems are mostly ignored, even though they severely limit the applicability of most word embedding algorithms as novel corpus linguistics methods. Unreliable methods can mislead users and contribute to the current reproducibility crisis (see Section 1.3). However, variants of the SVD_{PPMI} (Levy et al., 2015) word embedding algorithm—especially my novel $\text{SVD}_{\text{wPPMI}}$ (see Section 4.6)—to be perfectly reliable without any loss of accuracy.
This thesis also explores new ways to use information on word change to track changes in emotional connotation. It also describes the Jena Semantic Explorer (JeSemE; Hellrich & Hahn (2017b), Hellrich et al. (2018a)), a website giving non-technical users access to state-of-the-art distributional semantics. Prior work on historical lexical emotion is very limited and used simplistic emotion models (i.e., words are either 'positive' or 'negative'). My cooperation with Sven Buechel, however, lead to fine-grained analyses with a multi-dimensional model (Hellrich et al., 2019a). JeSemE is the first interactive website for accessing information on trends in both denotation and emotional connotation derived from word embeddings. It provides access to five

diachronic corpora in both German and English.
Finally, this thesis contains two case studies applying JeSemE to questions of interest to the digital humanities. Their results are in line with the scholarly expectations, indicating my approach to be reasonable.

1.5. Outline

Chapter 2 introduces distributional semantics in general and word embeddings in particular as the background for the experiments described in Chapters 4 and 5. Distributional judgments on word similarity match human intuition (Rubenstein & Goodenough, 1965), making them useful for further studies. Three popular word embedding methods, i.e., GloVe (Pennington et al., 2014), SGNS (Mikolov et al., 2013a,b) and SVD_{PPMI} (Levy et al., 2015), are discussed in detail. Chapter 2 also contains an overview of the history of distributional methods and their recent application to track lexical change.

Chapter 3 introduces several diachronic corpora which were used for the experiments in Chapters 4 and 5. It describes the composition of these corpora as well as peculiarities that might affect the interpretation of results in Chapter 5. This is especially important for the Google Books Ngram corpus (Lin et al., 2012; Michel et al., 2011), as its composition is opaque and seems to be unstable (Pechenick et al., 2015).

Chapter 4 investigates the (lack of) reliability of several word embedding algorithms introduced in Chapter 2. It describes a series of experiments focused on the SGNS algorithm and the effect of different sampling strategies applied during the training phase of word embeddings. It also contains a theoretical description of the source and possible consequences of the lack of reliability as well as a discussion of the—albeit scarce—related work.

Chapter 5 is concerned with the application of word embeddings in the digital humanities and linguistics. It contains experiments on using word embeddings to create a fine-grained model of past emotional connotation. Chapter 5 also introduces JeSemE, a website allowing non-technical users to profit from state-of-the-art

distributional semantics methods while avoiding reliability problems. This chapter also contains two case studies on applying JeSemE to investigate the history of electricity and the meaning of words connected to the period of Romanticism.

Chapter 6 summarizes the other chapters and provides general recommendations on best practices, possible applications and ideas for future research.

Chapter 2

Word Distribution and Meaning

Observing the frequency of words and word combinations to infer a word's meaning is an old idea at the heart of several state-of-the-art solutions in computational linguistics. It is intuitively appealing, since humans are able to learn the meaning of words from examples—in the most extreme case two persons can acquire a shared vocabulary without having any shared language, as in the following thought experiment (Wiener, 1955, p. 183):

> "Suppose I find myself in the woods with an intelligent savage, who cannot speak my language, and whose language I cannot speak. [...] [A] signal without an intrinsic content may acquire meaning in his mind by what he observes at the time, and may acquire meaning in my mind by what I observe at the time."

Language internal contexts, such as adjacent words,[1] can also be used to determine the meaning of a word, a concept known as distributional semantics (Harris, 1954; Rubenstein & Goodenough, 1965; Turney & Pantel, 2010) and succinctly described by the frequently cited "You shall know a word by the company it keeps!" (Firth, 1968, p. 11). Contextual information can be used to determine whether two words are especially likely to co-occur with each other and also whether they are similar—from a linguistic point of view, these are questions of syntagmatic and paradigmatic relations (Schütze & Pedersen, 1993). This corresponds to a structuralist line of linguistic theory described by Harris (1954, p. 157):

> "If A and B have some environments in common and some not (e.g. oculist and lawyer) we say that they have different meanings, the amount of meaning difference corresponding roughly to the amount of difference in their environments."

[1] Thus in this case a word's context is equal to its co-text.

A similar idea was expressed in the following sketch of a disambiguation algorithm by Weaver (1955, pp. 20–21):

> "If one examines the words in a book, one at a time as through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of the words. [...] But if one lengthens the slit in the opaque mask, until one can see not only the central word in question but also say N words on either side, then, if N is large enough one can unambiguously decide the meaning of the central word."

Progress in the fields of information theory[2] (Fano, 1966; Shannon, 1948) and information retrieval[3] (Deerwester et al., 1990; Salton, 1971) lead to modern algorithms for determining the meaning of words, especially the ways in which they are similar to each other. Distributional semantics are attractive for historical linguistic research, since text corpora are the main source for past language. In addition to diachronic applications, distributional methods can also be used to investigate other linguistic questions, e.g., language acquisition (Landauer & Dumais, 1997), regional variation (Hovy & Purschke, 2018; Kulkarni et al., 2016), changes affecting loan words (Takamura et al., 2017), or social biases reflected in language use (Bolukbasi et al., 2016a,b; Caliskan et al., 2017). They might also be useful for the digital humanities in general, e.g., Jo (2016) applied them to study diplomatic documents.

The following Section 2.1 describes different types of contexts, e.g., nearby words, and nuances involved in sampling texts. It also introduces the **word context matrix** which is instrumental to several methods in later Sections. Section 2.2 defines similarity and presents general methods for modeling word meaning. Section 2.3 introduces vector representation for word meaning, describes three popular word embedding algorithms in detail and provides a short overview of the history of these methods. Section 2.4 is concerned with the evaluation of methods for determining word similarity. Finally, Section 2.5 provides an overview of diachronic applications of distributional methods. Both overviews of distributional methods are impeded by

[2] Information theory was developed to provide mathematical underpinnings for signal processing and is related to statistics and stochastics.

[3] Information retrieval is mainly concerned with finding documents matching a user query, a widely used modern example are search engines, such as Google or Bing.

the relatively long research history of this currently very active field.[4] Examples for early applications or theoretical underpinnings are thus only intended to show the minimum age of individual approaches—a complete study of the history of applied distributional methods would involve archival research and probably constitute a thesis in itself.

2.1. Sampling Contexts

Contextual information for a word can be determined by a wide variety of approaches, e.g., the documents a word occurred in, grammatically dependent words, or co-occurring words (see e.g., Rubenstein & Goodenough (1965), Deerwester et al. (1990), Levy & Goldberg (2014a)). Words co-occurring in a small window (recall Weaver's algorithm sketch on page 10) were found to be an overall good choice for modeling lexical semantics (Levy et al., 2015; Mikolov et al., 2013b; Pennington et al., 2014) and are thus described here.
Examples in the remainder of this chapter are based on the following three sentence corpus:

Somebody buys a novel.
Somebody buys a book.
Somebody purchases a book.

Figure 2.1 shows a context window being moved over the first sentence of the example corpus (converted to lower case), sampling one co-occurring word to each side of the current word.

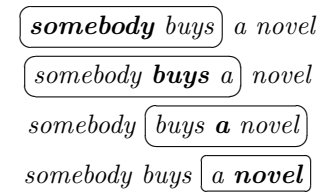

Figure 2.1: Visualization of a symmetrical one-word context window being moved over a sentence; current center words in **bold**.

[4] The most popular family of algorithms, i.e., `word2vec`, was cited over twenty three thousand times in the last six years. Checked for Mikolov et al. (2013a) and Mikolov et al. (2013b) on May 28th 2019 via `https://scholar.google.com/citations?user=oBu8kMMAAAAJ`

Window-based approaches necessitate the choice of window size (e.g., one in the above example) and a decision on symmetry—are words to both sides processed in the same way (symmetrical, as in the above example) or not (asymmetry). Another choice is using a bag-of-words approach and treat all contexts the same independently of their position or, alternatively, encoding positional information. Current research favors symmetric windows (Bullinaria & Levy, 2012) and a bag-of-words approach over the alternatives explored before (e.g., by Schütze (1993)). Recommended window sizes commonly vary from 2 to 5 (see e.g., Levy et al. (2015)). A now discarded alternative to these artificial windows are naturally defined windows, e.g., counting co-occurrences in document titles as in Lewis et al. (1967), documents as in Doyle (1961) or sentences as in Rubenstein & Goodenough (1965).

While some of the currently most popular methods operate in a streaming fashion and process one word context combination at a time, many utilize a **word context matrix**, M, which stores the connections between the i-th word and the j-th context in $M_{i,j}$. The word context matrix shown in Table 2.1 lists all words in the corpus presented on page 11 as processed with the method illustrated in Figure 2.1.

Contexts are often subjected to minimum frequency thresholds, e.g., only modeling co-occurrences with words that appeared at least 100 times. This kind of filtering is globally applied for all combinations of each context with any word. Another kind of filtering is locally applied only to some word context combinations, e.g., keeping only

	a	*book*	*buys*	*novel*	*purchases*	*somebody*
a	0	2	2	1	1	0
book	2	0	0	0	0	0
buys	2	0	0	0	0	2
novel	1	0	0	0	0	0
purchases	1	0	0	0	0	1
somebody	0	0	2	0	1	0

Table 2.1: Direct co-occurrence counts for words in the example corpus.

the 10 most frequent contexts for each word. Filtering can reduce noise as well as memory requirements (see page 21).
Another common modification of co-occurrence values applies some kind of downsampling (Levy et al., 2015; Mikolov et al., 2013b; Pennington et al., 2014). Downsampling can be used to reduce the impact of high frequency words (probably stop of function words) which would otherwise dominate the overall modeling efforts, a process called 'subsampling' in Levy et al. (2015).[5] Downsampling can also be used to reduce the impact of words that are relatively far from a modeled word (and thus assumed to be less relevant), a process called 'dynamic context window' in Levy et al. (2015). Both types of downsampling were shown to affect performance, with downsampling of high frequency words being especially helpful (Levy et al., 2015; Mikolov et al., 2013b)—see also Section 4.6.
Downsampling can be implemented as a probabilistic process or via weighting. In the former case co-occurrences are sampled with some probability, e.g., according to the distance between co-occurring words. In the latter case all co-occurrences are processed, but their impact, e.g., on counts in a word context matrix, is lowered by multiplication with a weight. Probabilistic approaches can be beneficial from a computational point of view, since only sampled instances need to be processed further in streaming algorithms. However, they contribute to the reliability problems described in Chapter 4. Details on downsampling procedures are given with each algorithm's description in Sections 2.3.2–2.3.4.

2.2. Word Association and Similarity

Early research did often not distinguish between association and similarity, or subsumed one under the other (Giuliano, 1963). Following Jurafsky & Martin (2009, ch. 20.7), association will here refer to syntagmatic patterns in the co-occurrence of a word with contexts. These co-occurrence patterns can be tracked either based on frequency, e.g., noun and articles will co-occur frequently, or based on expected and observed probabilities, e.g., *delicious* will likely co-occur with the names of dishes. Association is thus not to be understood in the sense of de Saussure's "rapports associatifs", which are decidedly

[5] Levy et al. (2015) also use 'context distribution smoothing' for a globally applied procedure with a similar goal.

non-syntagmatic and correspond to questions of similarity (Wunderli, 2013, pp. 262–263).
From a distributional point of view, words are similar if they share patterns in their associated contexts. They are thus paradigmatically related (see e.g. Landauer & Dumais (1997) or Bullinaria & Levy (2007)). Such distributional similarity has no direct connection to the outside world or classical linguistic semantics and might seem to be of limited use for investigating language change or (computational) linguistic applications in general. However, Rubenstein & Goodenough (1965) showed distributional similarity to be correlated with human similarity judgments, allowing for both applications (see especially Chapter 5) and evaluation (see Section 2.4).

2.2.1. Word Association Measures

Word association measures allow large corpora to be screened for typical examples of a word's usage or automatically identify multi-word expressions, e.g., the compound *Cold War*.[6] The former application is popular in corpus-based linguistic research in general (Biber et al., 2000; Curzan, 2009), while the latter can be used to create and curate terminological resources or templates for automatic text creation (e.g., Evert & Krenn (2001), Wermter & Hahn (2004), Smadja & McKeown (1990)). Early research on word association was motivated by the challenge of organizing large document collections and finding patterns for indexing or refining user queries about such collections (Doyle, 1961; Maron & Kuhns, 1960; Stiles, 1961).
Using an association measure to describe word combinations is qualitatively different from using the most frequent adjacent words as it is less likely to be affected by function words (Jurafsky & Martin, 2009, p. 695). There exists a wide range of word association measures, see e.g., Evert (2005) for an overview. Probably the most established word association measure is Pointwise Mutual Information (PMI), which was introduced by Fano (1966)[7] and popularized for word

[6] Highly associated after World War Two, but not before; see 'Typical Context' results for the `JeSemE` system described in Section 5.2: `http://jeseme.org/search?word=cold&corpus=coha` [Accessed May 28th 2019].

[7] Fano used the name 'mutual information', which nowadays refers to a more general measure which would describe the expected association for a randomly picked pair of words, i.e., something akin to text coherence or repetitiveness. See also Jurafsky & Martin (2009, p. 696).

association by Church & Hanks (1990). A variant that characterized words by the information provided by all their co-occurrences had already been employed earlier by Spiegel & Bennett (1965, p. 54). PMI is the logarithmized ratio of the observed co-occurrences between a word and a context (word) to the co-occurrences expected based on the independent occurrences of both. It can be calculated based on count-derived probabilities[8] for encountering the word i and context j alone, i.e., $P(i)$ and $P(j)$, and together, i.e., $P(i, j)$:

$$PMI(i, j) := log \frac{P(i, j)}{P(i)P(j)} \tag{2.1}$$

Using $PMI(buys,\ somebody)$ and $PMI(buys,\ a)$ as examples, the relevant probabilities according to Table 2.1 are:[9]

$$\begin{aligned}
P(a) &= \frac{6}{18} = \frac{1}{3} \\
P(somebody) &= \frac{3}{18} = \frac{1}{6} \\
P(buys) &= \frac{4}{18} = \frac{2}{9} \\
P(buys,\ somebody) &= \frac{2}{18} = \frac{1}{9} \\
P(buys,\ a) &= \frac{2}{18} = \frac{1}{9} \\
P(buys)P(a) &= \frac{2}{9} \times \frac{1}{3} = \frac{2}{27} \\
P(buys)P(somebody) &= \frac{2}{9} \times \frac{1}{6} = \frac{1}{27}
\end{aligned}$$

[8] For a Vocabulary V the probability of encountering a word $x \in V$ is calculated with the help of a function $\mathrm{c}(x)$, which provides the number of times a word x appeared in the corpus, and the following formula: $P(x) = \mathrm{c}(x)/\sum_{v \in V} \mathrm{c}(v)$. Note that this example did not distinguish between words and context words, as it is applied to symmetrical data.

[9] Probabilities derived directly from the example sentences would differ as the process of moving a context window over the text inflates and distorts counts. For example, each of the three sentences contains an a and three other words, thus $P(a) = 1/4$ when derived directly from the corpus.

Assuming base 2 for the logarithm[10] the resulting PMI values are:

$$PMI(buys,\ somebody) = log(\frac{1/9}{1/27}) = log(3) \approx 0.48$$
$$PMI(buys,\ a) = log(\frac{1/9}{2/27}) = log(1.5) \approx 0.18$$

According to these PMI values *buy* is more strongly associated with *somebody* than with *a*. While both co-occurrences are equally frequent, *a* co-occurs with a greater number of different words and is thus less specific.
Pearson's χ^2 which is also used as a statistical test for the association between categorical variables (e.g., parts of speech), is a robust alternative to PMI (Manning & Schütze, 1999, ch. 5). An early linguistic application was provided by Stiles (1961) identifying synonyms and other closely related words among terms used for indexing with a variant of χ^2. The χ^2 association[11] between a word i and a context j is also calculated with count-derived probabilities:

$$\chi^2(i,j) := \frac{(P(i,j) - P(i)P(j))^2}{P(i)P(j)} \tag{2.2}$$

χ^2 again indicates *buys* to be more associated with *somebody* than with *a*:

$$\chi^2(buys,\ somebody) = \frac{(1/9 - 1/27)^2}{1/27} \approx 0.15$$
$$\chi^2(buys,\ a) = \frac{(1/9 - 2/27)^2}{2/27} \approx 0.02$$

Positive Pointwise Mutual Information (PPMI) is a variant of PMI independently developed by Niwa & Nitta (1994) and Bullinaria & Levy (2007). It provides only positive values which indicate words to co-occur more often than expected by chance:

$$PPMI(i,j) := \begin{cases} 0 & \text{if } \frac{P(i,j)}{P(i)P(j)} < 1 \\ log(\frac{P(i,j)}{P(i)P(j)}) & \text{otherwise} \end{cases} \tag{2.3}$$

[10] The choice of base is irrelevant for comparisons between different words.
[11] All of these $\chi^2(i,j)$ values would be combined for a χ^2 test.

This achieves not only better results but is also mathematically beneficial, since PMI itself is not defined for word combinations that never occurred as it would require calculating $log(0)$.
Note that $PPMI(i,j) = 0$ for words that never co-occurred, while $\chi^2(i,j) \neq 0$ for nearly all word combinations. This is problematic, as it leads to increased memory consumption, despite both measures having an asymptotic space complexity of $\mathcal{O}(n^2)$ for a vocabulary of size n. However, in practice matrices for PPMI contain about 99% zeroes which can be utilized to save memory through sparse matrix formats (see discussion on page 21). Later experiments in this thesis use a sparse version of Equation 2.2:

$$\chi^2(i,j) := \begin{cases} 0 & \text{if } P(i,j) = 0 \\ \frac{(P(i,j)-P(i)P(j))^2}{P(i)P(j)} & \text{otherwise} \end{cases} \tag{2.4}$$

Word association measures are evaluated by their ability to solve a task, e.g., identifying collocations. This can be done both by manually inspecting the most associated words and also by comparison with a gold standard. The latter requires a higher initial time investment, yet allows for cheap future evaluations and the calculation of recall scores, i.e., measuring missed potential matches (Evert & Krenn, 2001).

2.2.2. Word Similarity

Some researchers distinguish between two types of similarity, hereinafter referred to as 'strict similarity' and 'relatedness' (Hill et al., 2014). Strict similarity assumes that the referents of words share attributes, e.g., *dog* is similar to *wolf* since both animals share much of their behavior and anatomy. Words with enough overlap are synonyms and can be used interchangeably, e.g., *to buy* and *to purchase*. Meanwhile relatedness arises from words being used in similar communicative situations, e.g., *dog* being similar to *cat* and even *kibble* due to people writing or talking about pets.
An alternative definition for strict similarity and relatedness is subsuming the former under the latter, with similar words being connected through "hyponymy (hypernymy), antonymy, or troponymy" (Mohammad, 2008, p. 3) while related words can be connected through any kind of semantic relationship, e.g., meronymy or both being positive adjectives (Mohammad, 2008, p. 2).

Both definitions suffer from the ontological problem of nearly all words being connected by trivial attributes like 'exists', respectively trivial hypernyms like 'thing' or 'action'. The attribute based definition can also be argued to be not well defined inside linguistics, since it depends on world knowledge.

The experiments described in Chapter 5 operate on a relatedness based definition, both due to most data sets and methods being intended for it and due to an assumed fit between its somewhat fuzzy nature and observing semantic change. They thus use co-occurrence derived contexts instead of syntactical ones, the latter being known to favor strict similarity (see e.g., Turney & Pantel (2010), Levy & Goldberg (2014a)).

Early research in word similarity measures was motivated by the challenge of managing large document collections (Giuliano, 1965; Lewis et al., 1967). One notable exception from this applied research is Rubenstein & Goodenough (1965) testing the linguistic distributional hypothesis by comparing a corpus based word similarity measure with human judgments.

A still common use case for word similarity measures is the construction of so called distributional thesauri (Ferret, 2017; Grefenstette, 1994; Salton & Lesk, 1971). These thesauri can serve as a resource in technical applications, especially so in information retrieval (Salton & Lesk, 1971). They enable systems to find not only verbatim matches, but also matches based on synonymy or hypernymy, e.g., documents containing *car* for a query concerned with *vehicle*. This information retrieval driven research led to the vector space model described in the next section.

Current methods do typically not use straight co-occurrence frequency to model similarity, but rely on some form of association measure (see Section 2.2.1) calculated in a separate step (Bordag, 2008). This measure can then be used in place of frequencies for calculating a similarity score (Lin, 1998) or as a filtering criterion (see discussion on page 21).

While most of the filtering approaches fall under the vector space concept and are thus described in the next section, some operate on a notion of sets, i.e., use only the information that two words were associated, but not to what degree (as long as it exceeds a minimum). The earliest example is Rubenstein & Goodenough (1965) who quantified the similarity of two words as the normalized number

of contexts both co-occurred with. A recent example is Riedl & Biemann (2013) representing words with sets containing their most associated grammatically dependent contexts.

2.3. Words as Vectors

The vector space model is now widely used for both distributional semantics and information retrieval (Salton et al., 1975; Turney & Pantel, 2010). It represents words with coordinates, allowing for comparisons between words through their positions. Such a geometric approach was already described for document indexing by Maron & Kuhns (1960, pp. 224–225):

> "The points in this space are not located at random, but rather, they have definite relationships with respect to one another, depending on the meanings of the terms. For example, the term 'logic' would be much closer to 'mathematics' than to 'music'."

In the simplest case each dimension or axis represents one context (e.g., *a* or *somebody*) and the coordinate on this axis represents the frequency or association with this context. Figure 2.2 shows the positions of *book*, *buys* and *purchases* in such a vector space (according to counts in Table 2.1). All vectors start from the origin and only axes for co-occurrences with *a* and *somebody* are shown.

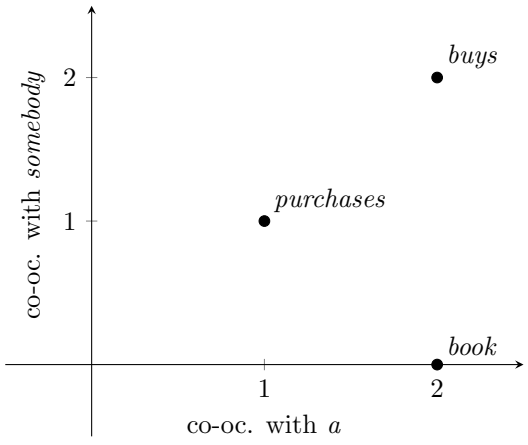

Figure 2.2: Positions of *book*, *buys* and *purchases* in a vector space with axes for co-occurences according to Table 2.1.

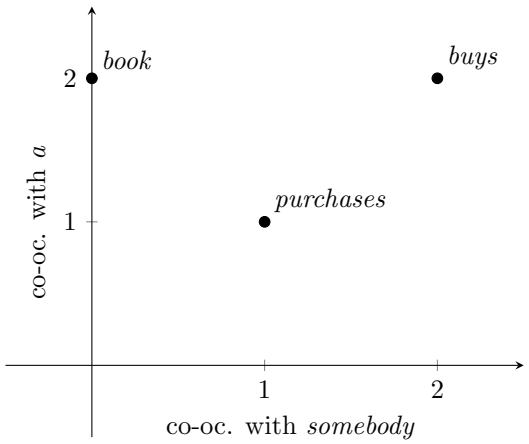

Figure 2.3: Rotation of the vector space from Figure 2.2.

Note that there is no inherent connection between a specific axis and the context it is used to track, i.e., data can be rotated by switching axes while preserving relative positions, as shown in Figure 2.3.

Word embeddings use fewer dimensions to represent each word in a vector space than contexts exist.[12] For example, Table 2.1 can be argued to contain several redundant entries. The words *book* and *novel*, respectively *buys* and *purchases*, are identically distributed, i.e., the columns describing their context words are identical.

Both these identical distributions, as well as linguistic intuition—*buys* and *purchases* are synonyms, whereas *book* is the hypernym of *novel*—make it plausible to merge these columns, resulting in Table 2.2. This manual approach towards dimensionality reduction results in interpretable dimensions, e.g., *buys & purchases*, whereas the algorithms described in the following sections produce opaque dimensions.

Word embeddings are not the only options to achieve more memory efficient representations. Set approaches (see page 18) or aggressive filtering (reducing most contexts to zero) can lead to very similar

[12] In principle every representation of a word in vector space could be called a word embedding, but the term is strongly associated with low dimensional representations (Turian et al., 2010, p. 386).

	a	*book & novel*	*buys & purchases*	*somebody*
a	0	2	2	0
book	1	0	0	0
buys	1	0	0	1
novel	1	0	0	0
purchases	1	0	0	1
somebody	0	0	2	0

Table 2.2: Co-occurrence counts for words in example corpus after manual dimensionality reduction; based on Table 2.1.

decreases in space complexity given a proper data structure.[13] Representation of the top d individual contexts of a word (including set representations), as well as word embeddings of dimensionality d for a vocabulary of size n have a space complexity of $\mathcal{O}(n \times d)$, whereas a non-sparse $word \times context\ word$ matrix has a complexity of $\mathcal{O}(n \times n)$ and typically $n \gg d$.

In contrast to other approaches, word embeddings are opaque, i.e., their dimensions do not directly correspond to contexts in the underlying data. Such opaque dimensions can be argued to be acceptable from a linguistic point of view, since at least synonymous context words can be exchanged without a (major) loss of information. Due to their usefulness in applications and high performance in judging word similarity[14] (see e.g., Levy et al. (2015) or Sahlgren & Lenci (2016)), word embeddings are currently extremely popular in computational linguistics, especially so the `word2vec` algorithms (Mikolov et al., 2013a,b).

[13] Suitable choices for non-word embedding approaches are associative data structures (e.g., used by Gamallo & Bordag (2011)) as well as the sparse matrices included in most numerical software (e.g., in MATLAB, see Gilbert et al. (1992)).

[14] The only recent reports of achieving superior performance with non-word embedding were made by Gamallo et al., who argue against word embeddings due to their opaqueness. They used the cosine (see below) between vectors containing only the top grammatical contexts by association score (Gamallo, 2017; Gamallo et al., 2018). Levy et al. (2015) provided good, but not quite state-of-the-art results with a similar approach, where all association scores were decreased by a constant factor, retaining only positive ones.

Sections 2.3.2–2.3.4 introduce three popular algorithms for creating word embeddings. Some of these approaches operate on initial high-dimensional representations, which are then transformed in a dimensionality reduced representation, while others avoid any form of high dimensional representation. These algorithms are also used in later chapters as they are popular and perform well. Niche alternative approaches like random indexing (Kanerva et al., 2000) and Hellinger-PCA (Lebret & Collobert, 2015) are thus out of scope. A general overview of the history of vector representations for words is given in Section 2.3.5.

2.3.1. Interpreting Vector Spaces

The relative positions of words in a vector space can be used for comparisons as in the quote from Maron & Kuhns (1960) on page 19. This can be done by calculating some kind of distance between words (see e.g., Bullinaria & Levy (2007)), as shown in Figure 2.4. Euclidean distance d conforms to an intuitive concept of distance in space and is calculated for two vectors a and b with n entries with:

$$d(a,b) := \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \tag{2.5}$$

An alternative is using the angle θ between two vectors, also shown in Figure 2.4. Empirical investigations overall show the angle, expressed as the cosine, to be superior during evaluation tasks (Bullinaria & Levy, 2007), but other measures might be beneficial for specific tasks or word frequency bands (Weeds et al., 2004).
The cosine typically ranges between 0 for complete dis-similarity and 1 for complete similarity.[15] As an angular measurement the cosine is only concerned with vector direction and not with vector magnitude. Cosine similarity is calculated for two vectors a and b with:

$$cos(a,b) := \frac{a \cdot b}{\|a\|\|b\|} \tag{2.6}$$

$a \cdot b$ calculates a vector dot product between a and b, i.e., the sum of the products of all corresponding n components:

$$a \cdot b := \sum_{i=1}^{n} a_i b_i \tag{2.7}$$

[15] Its complete range is from -1 to 1. Negative values do not correspond to linguistic antonymy. In fact, distributional models tend to rate antonyms as rather similar and seldom provide negative cosine values.

The magnitude of a vector x of length n is defined as:

$$\|x\| := \sqrt{\sum_{i=1}^{n} x_i^2} \tag{2.8}$$

For example, the cosine between *buys* and *purchases* represented by vectors b and p (for *buys* and *purchases*, respectively) can be calculated for Table 2.1 with:

$$\begin{aligned} b &= [2\ 0\ 0\ 0\ 0\ 2] \\ p &= [1\ 0\ 0\ 0\ 0\ 1] \\ cos(b,p) &= \frac{[2\ 0\ 0\ 0\ 0\ 2] \cdot [1\ 0\ 0\ 0\ 0\ 1]}{\|[2\ 0\ 0\ 0\ 0\ 2]\|\ \|[1\ 0\ 0\ 0\ 0\ 1]\|} \\ &= \frac{4}{\sqrt{8}\sqrt{2}} \\ &= 1 \end{aligned}$$

The empirical superiority of the cosine might be due to its resilience against variations in absolute word frequency (due to the division by the magnitudes), relying on relative frequencies instead. As highlighted in Figure 2.4, *purchases* has the same distance to *buys* and *book*, but widely different angles (the 0° angle between *purchases* and *buys* not being visible). Only the latter reflects *buys* occurring with the same words as *purchases*.

Operations in vector space can not only be used to measure word similarity, but also to solve analogy tasks (Mikolov et al., 2013a), e.g., which word relates to *king* as *woman* does to *man*. Assume four word embeddings a, b, c, d (all elements of a vocabulary V) where a (e.g., *man*) and b (e.g., *woman*) are to each other as c (e.g., *king*) is to an unknown d. Mikolov et al. (2013a) identified d by maximizing the following Equation:

$$\underset{d \in V}{arg\,max}(cos(d,\ c - a + b)) \tag{2.9}$$

Equation 2.9 can be rewritten as (Levy & Goldberg, 2014b):

$$\underset{d \in V}{arg\,max}(cos(d,c) - cos(d,a) + cos(d,b)) \tag{2.10}$$

Intuitively, Equation 2.9 describes a movement from c towards b (which share some property with d) and away from a (which is lacking

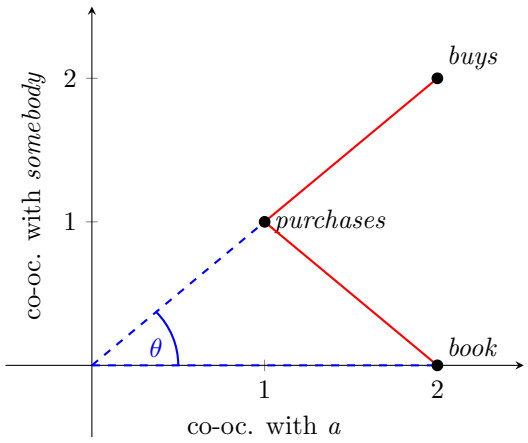

Figure 2.4: Comparisson of *purchases* to *book* and *buys* by distance, as well as *book* by angle; based on Figure 2.2.

this property). This corresponds, as highlighted in Equation 2.10, to searching for a d close to c and b, but far from a.
Levy & Goldberg (2014b) found results of this additive approach to be strongly influenced by the largest difference between terms. They achieved a better balance between large and small differences by effectively "taking the logarithm of each term before summation" (Levy & Goldberg, 2014b, p. 175). This multiplicative approach also necessitates the addition of a small ϵ to prevent division by zero:

$$\underset{d \in V}{arg\,max}\left(\frac{cos(d,c)\ cos(d,b)}{cos(d,a)+\epsilon}\right) \tag{2.11}$$

The ability to solve analogies can be used to evaluate algorithms or parameter choices. Mikolov et al. (2013c) created a test set with 8,000 morpho-syntactic analogy questions (e.g., *year* is to *years* as *law* is to *?*) at Microsoft Research.[16] Mikolov et al. (2013a) also created a test set with 19,544 questions of which about half are morhpo-syntactic (e.g., *think* is to *thinking* as *read* is to *?*) and half are semantic[17] (e.g., *Athens* is to *Greece* as *Oslo* is to *?*) at Google.

[16] The provenance is given, as both test sets are often referenced by it, e.g., 'MSR' and 'Google' in Levy & Goldberg (2014b).

[17] This example can be argued to test encyclopedic knowlegde from a linguistic point of view.

There is a fundamental difference between vector representations of words and thesauri—the former are bound by geometrical constraints, while the latter are not. Thus for three words w_1, w_2 and w_3 the two similarity values $sim_1(w_1, w_2)$ and $sim_2(w_2, w_3)$ would constrain the possible value of $sim_3(w_1, w_3)$ in vector space, like two sides of a triangle constrain the third. This is not the case for a thesaurus, which can be modeled as a graph with weighted edges, allowing for arbitrary similarity values between w_1, w_2 and w_3.

2.3.2. SVD_{PPMI} Word Embeddings

Singular Value Decomposition (SVD) can be used to automatically decrease the dimensionality of a word context matrix and thus create word embeddings. The state-of-the-art method for creating word embeddings with SVD operates on a word context matrix pre-processed to contain Positive Pointwise Mutual Information (SVD_{PPMI}; Levy et al. (2015)).
In general, SVD represents a matrix M as the product of three special matrices (Berry, 1992; Saad, 2003):

$$M = U\Sigma V^T \tag{2.12}$$

Here U and V are orthogonal matrices containing so called singular vectors. Σ is a diagonal matrix containing singular values.[18] For a matrix M of rank r all diagonal entries $\sigma_i > 0$ for $1 \leq i \leq r$ and $\sigma_i = 0$ for $i > r$. The singular values are typically sorted in decreasing order, i.e., $\sigma_i \geq \sigma_{i+1}$.[19] The size of each singular value σ_i can be interpreted as the importance of the corresponding vectors in U and V.
Table 2.1 can be expressed with the following matrices, rounding everything to two significant digits for readability:[20]

[18] Singular values are frequently called eigenvalues in literature, e.g., in Levy et al. (2015). This is due to the Eigendecomposition procedure, which is roughly speaking a variant of SVD suitable only for square matrices.

[19] At least in the case of the LAS2 algorithm by Berry (1992) used in Levy et al. (2015); other orders would require all matrices to be rearranged for further processing.

[20] Calculated with the SVD implementation of `NumPy` [Accessed May 28th 2019]: `https://docs.scipy.org/doc/numpy-1.13.0/reference/generated/numpy.linalg.svd.html`

$$U \cong \begin{bmatrix} 0.85 & 0 & 0.53 & 0 & 0 & 0 \\ 0 & -0.37 & 0 & -0.60 & 0.38 & 0.60 \\ 0 & -0.6 & 0 & 0.37 & 0.60 & -0.38 \\ 0 & -0.37 & 0 & -0.60 & -0.38 & -0.60 \\ 0 & -0.60 & 0 & 0.37 & -0.60 & 0.37 \\ 0.53 & 0 & -0.85 & 0 & 0 & 0 \end{bmatrix}$$

$$\Sigma \cong \begin{bmatrix} 2.29 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2.29 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.87 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.87 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$V^T \cong \begin{bmatrix} 0 & 0.37 & 0.60 & 0.37 & 0.60 & 0 \\ -0.85 & 0 & 0 & 0 & 0 & -0.53 \\ 0 & 0.60 & -0.37 & 0.60 & -0.37 & 0 \\ -0.53 & 0 & 0 & 0 & 0 & 0.85 \\ 0 & -0.11 & 0.70 & 0.11 & -0.70 & 0 \\ 0 & 0.70 & 0.11 & -0.70 & -0.11 & 0 \end{bmatrix}$$

$$U\Sigma V^T \cong \begin{bmatrix} 0.00 & 1.00 & 1.00 & 1.00 & 1.00 & 0.00 \\ 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.01 \\ 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \\ 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.01 \\ 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \\ 0 & 0.01 & 1.00 & 0.01 & 1.00 & 0.00 \end{bmatrix}$$

Dimensionality reduction can be achieved by using only the top d entries of Σ and the corresponding singular vectors in U_d and V_d. The result of this so called economical SVD is a matrix M_d which is similar, but not identical, to M:

$$M_d = U_d \Sigma_d V_d^T \tag{2.13}$$

For example, using only 3 dimensions and rounding to two significant digits results in M_3 as an imperfect reconstruction of Table 2.1:

$$U_3 \cong \begin{bmatrix} 0.85 & 0 & 0.53 \\ 0 & -0.37 & 0 \\ 0 & -0.60 & 0 \\ 0 & -0.37 & 0 \\ 0 & -0.60 & 0 \\ 0.53 & 0 & -0.85 \end{bmatrix}$$

$$\Sigma_3 \cong \begin{bmatrix} 2.29 & 0 & 0 \\ 0 & 2.29 & 0 \\ 0 & 0 & 0.87 \end{bmatrix}$$

$$V_3^T \cong \begin{bmatrix} 0 & 0.37 & 0.60 & 0.37 & 0.60 & 0 \\ -0.85 & 0 & 0 & 0 & 0 & -0.53 \\ 0 & 0.60 & -0.37 & 0.60 & -0.37 & 0 \end{bmatrix}$$

$$M_3 = U_3\Sigma_3V_3^T \cong \begin{bmatrix} 0.00 & 1.00 & 1.00 & 1.00 & 0.1.00 & 0.00 \\ 0.72 & 0.00 & 0.00 & 0.00 & 0.00 & 0.45 \\ 1.17 & 0.00 & 0.00 & 0.00 & 0.00 & 0.73 \\ 0.73 & 0.00 & 0.00 & 0.00 & 0.00 & 0.45 \\ 1.17 & 0.00 & 0.00 & 0.00 & 0.00 & 0.73 \\ 0.00 & 0.01 & 1.00 & 0.01 & 1.00 & 0.00 \end{bmatrix}$$

SVD_{PPMI} applies economical SVD to a matrix of PPMI values (see Equation 2.3). The vectors in U_d are then used as word representations and those in V_d as context word representations. In contrast to older SVD based word embedding approaches (e.g., Bullinaria & Levy (2007)), Σ is only used for dimensionality reduction and not used to scale U.

Co-occurrence counts in SVD_{PPMI} are downsampled both by distance between word and context word and by word frequency. The distance between two tokens w_i and w_j (the latter here serving as context for modeling the former) can be calculated for a sequence of tokens $\ldots, w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}, \ldots$ with:

$$d(w_i, w_j) := |j - i| \tag{2.14}$$

Downsampling by distance is performed by applying a window factor wf_{SVD} to each individual co-occurrence for a window of size s:

$$wf_{SVD}(w_i, w_j) := \frac{s + 1 - d(w_i, w_j)}{s} \tag{2.15}$$

Word frequency is downsampled with two strategies, one applied to all words and one only to high frequency ones. All contexts are downsampled during PMI calculation by modifying $P(j)$ in Equation 2.1 and raising the counts (provided by $c(v)$) by α (typically $\alpha = 0.75$):

$$P(j) := \frac{c(j)^\alpha}{\sum_{v=1}^{V} c(v)^\alpha}$$

Especially frequent words are further downsampled with a frequency factor ff_{SVD} based on a threshold t (typically $t = 10^{-5}$ or 10^{-4}) and each word's relative frequency provided by $r(w)$:

$$ff_{SVD}(w) := \begin{cases} \sqrt{t/r(w)} & \text{if } r(w) > t \\ 1 & \textit{otherwise} \end{cases} \tag{2.16}$$

Frequency downsampling for the co-occurrences of two tokens w_i and w_j is given by the product of their frequency factors, i.e., co-occurrences are treated as independent events:

$$ff_{SVD}(w_i, w_j) := ff_{SVD}(w_i)\, ff_{SVD}(w_j)$$

Levy et al. defined the downsampling factors wf_{SVD} and ff_{SVD} as optional and probabilistic, i.e., as a chance to sample a co-occurrence. Experiments in Section 4.6 showed a novel variant using weighting-based sub-sampling, i.e., using wf_{SVD} and ff_{SVD} as weights while populating M, to be perfectly reliable without sacrificing accuracy—we coined this variant $\text{SVD}_{\text{wPPMI}}$ (SVD based on a PPMI matrix populated via **weighting**).

SVD always provides the same matrix decomposition for a given input as long as no probabilistic downsampling is performed. This is so despite a random vector being used as a kind of anchor during the SVD calculation (Saad, 2003, chs. 6.3 & 7.1). An exception is stochastic SVD (Halko et al., 2011) which generates non-identical embeddings during repeated calculation (Antoniak & Mimno, 2018). Stochastic SVD is beneficial for processing streamed input as it requires only a single pass over the data and can be adapted not to require the pre-computation of a co-occurrence matrix.

2.3.3. Skip-gram Word Embeddings

Skip-gram (SG) word embeddings (Mikolov et al., 2013a,b) were shown to be superior to older SVD approaches on a wide variety of tasks (Baroni et al., 2014). They are not derived from a co-occurrence matrix through dimension reduction, but are initialized with random values and then tuned to predict likely context words given a word in question. This is done with an artificial neural network (NN), a machine learning method inspired by the interconnection of biological neurons and currently very popular in the form of deep learning (Goodfellow et al., 2016; LeCun et al., 2015). For both approaches, contexts are downsampled with a probabilistic window and frequency factor. In contrast to SVD_{PPMI}, they process texts in a streaming fashion and do not require the creation of a word context matrix.
Typical NNs consist of neurons arranged into one input layer, an arbitrary number of so called hidden layers and one output layer. The input of each neuron on one layer is determined by the output of all neurons on the previous layer.[21] Connections between these layers are described by matrices as well as (non-linear) activation functions. Mikolov et al. (2013a) introduced two architectures: Continuous Bag-of-Words (CBOW) for predicting a center word given context words and Skip-gram (SG) for predicting context words given a center word. Both are also known as `word2vec`, the name of the tool providing their reference implementations.[22] In general, SG embeddings are superior (see e.g., Levy et al. (2015)) and were thus used for most of the experiments described in Chapter 4.
The architecture of a SG network has three layers and is illustrated in Figure 2.5. It consists of an input layer l^I with one neuron per word in the vocabulary[23] to allow for one-hot encoding, i.e., each word is encoded by exactly one neuron being 1 and all others 0. For example, a vocabulary of size $V = 3$ requires an input layer with 3 neurons and would encode words with the one hot row vectors [1 0 0], [0 1 0] and [0 0 1]. The hidden layer l^H has one neuron per intended dimension d of the resulting embeddings. The output layer l^O is analogous to the input layer, again encoding each word in the

[21] Some architectures allow for other types of links, e.g., skipping layers.
[22] `https://github.com/tmikolov/word2vec` [Accessed May 28th 2019].
[23] The following text assumes words and context words to be from the same vocabulary, but formulas can easily be adjusted for differing vocabularies, e.g., due to syntactic contexts being used.

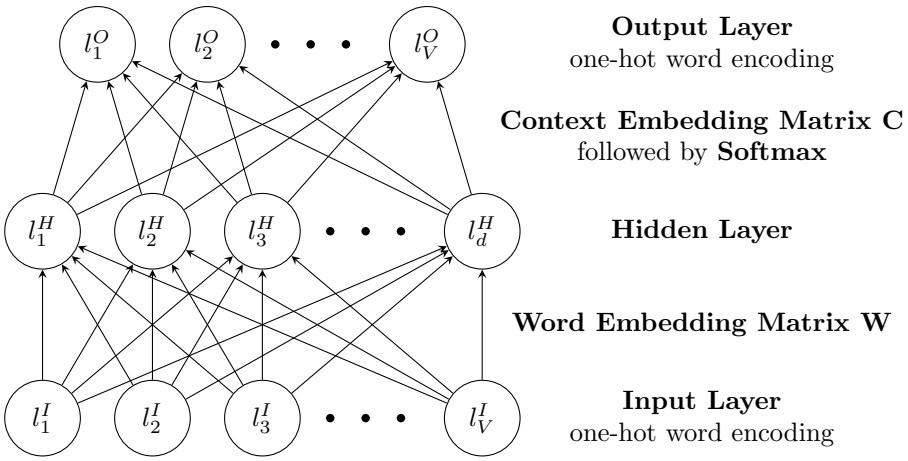

Figure 2.5: Skip-gram embedding NN architecture.

vocabulary with a one hot vector. The connection between l^I and l^H consist of a word embedding matrix $W_{V\times d}$ applied without any further function, i.e., $l^H = l^I W$, thereby selecting the word embedding for a given input word. l^H and l^O are connected by a context embedding matrix $C_{d\times V}$ and the softmax function, i.e., $l^O = \text{softmax}(l^H C)$. Softmax is commonly used for NN classification tasks (Goodfellow et al., 2016, pp. 180–184) and is used here to reconstruct a one-hot vector indicating a context word. It transforms arbitrary real valued vectors to vectors representing probabilities, i.e., their components (all between 0 and 1) sum up to 1. See Equation 2.18 for details.

During training actual and desired output of a neural network are compared and the entries of their weight matrices updated with a technique called backpropagation (see e.g., Goodfellow et al. (2016, ch. 6.5)) to reduce aberrations, starting from the output layer and propagating backwards to the input layer. This is done with a stochastic gradient descent algorithm, i.e., vector entries are modified in such a way that the overall error is minimized by following the slope of a function (see e.g., Goodfellow et al. (2016, ch. 5.9)), as illustrated in Figure 2.6. Stochastic gradient descent does not result in a global minimum, but will find one of multiple local minima instead. Note that the updates provided by the gradient descent algorithm are not

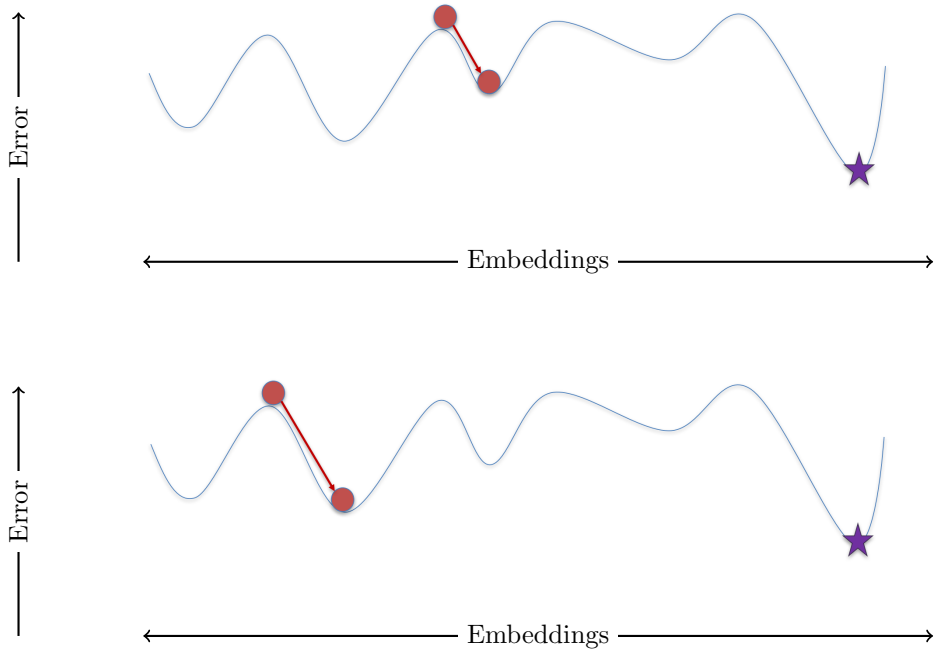

Figure 2.6: Illustration of the influence of starting positions on the selection of local minima (in red) during stochastic gradient descent. Combined effect of all embedding dimensions shown as one axis. Global maximum in purple.

applied fully, but multiplied with a factor known as learning rate (e.g., 0.025, typically decreased during ongoing training) to moderate the impact of each processed example.
The choice of a function for measuring such errors and guiding the gradient descent, called a loss function, is another important choice when designing a neural network. Following Mikolov et al. (2013b), SG tries to minimize false predictions for the s words before and after each word in a sequence of T words $w_1, w_2, \ldots w_T$:

$$error_{SG} := -\frac{1}{T}\sum_{t=1}^{T}\sum_{-s\leq j\leq s,\, j\neq 0} log\ p(w_{t+j}|w_t) \tag{2.17}$$

The conditional probability for encountering a context word c given an input word w is defined with the following softmax function:

$$p(c|w) := \frac{exp(C^T_{*,c} \cdot W_{w,*})}{\sum_{v=1}^{V} exp(C^T_{*,v} \cdot W_{w,*})} \tag{2.18}$$

Here $W_{w,*}$ refers to the row in W representing w, while $C_{*,c}$ refers to the column in C representing c and $C_{*,v}$ to the vth column in C (used to address vectors corresponding to all members of a vocabulary of size V).[24] After training, Mikolov et al. use the rows of W as word representations. It is possible, but without a clear benefit, to use the sum of corresponding rows and columns in both matrices instead (Levy et al., 2015), i.e., $W_{v,*} + C^T_{*,v}$ for a word v.

This basic Skip-gram algorithm is very unpractical, since it requires a comparison with all V words in the vocabulary due to the divisor in Equation 2.18. Mikolov et al. (2013b) offer two solutions: Skip-gram Hierarchical Softmax (SGHS) and Skip-gram Negative Sampling (SGNS). SGHS uses a binary tree to reduce the number of comparisons to $log\ _2(V) - 1$. Leaves (nodes without children) encode words while inner nodes (with children) represent ways towards those. Thus the task becomes one of learning the correct path through the inner nodes that need to be passed from the root to the correct leaf for a context word. In contrast, SGNS simply draws n random words from the vocabulary,[25] typically $n = 5$, and uses those instead of all V vocabulary entries during softmax calculation. Both SGNS and SGHS require more complex loss calculations (see Mikolov et al. (2013b)).

Downsampling factors for nearby words and high frequency words in SGNS embeddings are defined as in SVD_{PPMI} embeddings,[26] thus $wf_{SGNS} = wf_{SVD}$ and $ff_{SGNS} = ff_{SVD}$. Mikolov et al. implemented both factors probabilistically, i.e., words are sampled according to these factors. Such probabilistic downsampling decreases training time as embedding updates are only calculated for sampled word context

[24] These connections between vectors and rows or columns are given by the definitions of l_H and l_O and the way in which row vector matrix products work. Some vectors are transposed due to the dot product definition.

[25] The chance of drawing a word is set to its relative frequency modified by an exponent α, typically $\alpha = 0.75$.

[26] However, Levy et al. (2015, pp. 214–215) found a deviation in `word2vec`'s implementation of *ff*.

combinations.[27] Mikolov et al. provided an option not to downsample high frequency words, but found it to decrease model quality (Mikolov et al., 2013b). Levy et al. (2015) provided an implementation without wf_{SGNS} which they found to be of equal performance. Section 4.6 describes experiments with a novel variant of SGNS that processes all examples and uses weighting-based downsampling, i.e., gradient updates are multiplied with the appropriate *wf* and *ff*.

2.3.4. GloVe Word Embeddings

Another popular algorithm, GloVe (Global Vectors; Pennington et al. (2014)), can be seen as somewhat of a hybrid between the previous two approaches—a pre-computed co-occurrence matrix is used as in SVD_{PPMI}, but word vectors are randomly initialized and tuned with stochastic gradient descent as in the `word2vec` algorithms. Instead of piecemeal processing the input text as the `word2vec` algorithms, the non-zero entries of the co-occurrence matrix are processed in random order. GloVe is based on the idea that similar words should have similar ratios between their co-occurrence frequencies with other words (used explicitly) and their own frequency (captured in bias terms in the final formula). Avoiding the non-linear activation functions in NN was a design goal of `GloVe` (Pennington et al., 2014, p. 1534).[28]
Its loss function is:

$$error_{\texttt{GloVe}} := \sum_{i,j=1}^{V} f(M_{i,j})(W_{i,*} \cdot C^T_{*,j} + b_i + b'_j - log(M_{i,j}))^2 \quad (2.19)$$

$W_{i,*}$ and $C_{*,j}$ are defined as for Skip-gram embeddings, i.e., rows and columns of matrices for words and context words, respectively. V is again the size of the vocabulary, while b_i and b'_j are bias terms which serve to model the frequency of the (context) word in question. Entries of $M_{i,j}$ can be weighted by a function $f(x)$, which provides a

[27] For example, distance based downsampling reduces the number of processed examples by 40% for a window of size 5. In general, only $\frac{s+1}{2s}$ instead of $2s$ need to be processed for a symmetric window of size s.

[28] However, frequency weighting is non-linear, see Equation 2.3.4.

mix of context distribution smoothing (recommended $\alpha = 0.75$) and subsampling of frequent words (recommended $x_{max} = 100$):[29]

$$f(x) := \begin{cases} (x/x_{max})^{\alpha} & \text{if } x < x_{max} \\ 1 & \text{otherwise} \end{cases}$$

Pennington et al. (2014) describe `GloVe` with word representations composed by adding corresponding entries of both matrices, i.e., $W_{v,*} + C^T_{*,v}$ for a word with index v in both. This results in something akin to an ensemble model, but a later survey found no clear benefit over using $W_{v,*}$ alone (Levy et al., 2015). In contrast to SG no further tuning is necessary to speed up training, since f(0) = 0, i.e., calculations are necessary only for those words that actually co-occurred (Pennington et al., 2014, p. 1536). The overall performance of `GloVe` in comparison with `word2vec` is disputed and likely lower (Levy et al., 2015; Pennington et al., 2014).
Downsampling based on distance inside the context window is determined by the following window factor:

$$wf_{\text{GloVe}}(w_i, w_j) := \frac{1}{d(w_i, w_j)} \tag{2.20}$$

Downsampling for GloVe is canonically implemented via weighting.

2.3.5. History and Research Trends

Despite their recent popularity, word embeddings are a surprisingly old approach (see e.g., Sahlgren (2006), Turney & Pantel (2010), Clark (2015)). The earliest proto-word vectors stem from psychological research on word association (Osgood, 1952, 1953; Osgood et al., 1957). Osgood et al. found patterns which were stable between different participants and could be interpreted as some kind of semantic space. However, these representations are more connected with word emotion lexicons (see Section 5.1) and their manual creation makes them a non-distributional method.
Word vectors proper (and arguably also word embeddings) stem from information retrieval research and the development of the vector space

[29] Frequency downsampling in GloVe is thus based on the conditional probability for two words co-occurring, whereas SVD_{PPMI} and SG use the product of two independent probabilities.

model for documents (Salton et al., 1975; Turney & Pantel, 2010). Giuliano & Jones (1962) (as well as Giuliano (1963,1965)) modeled word similarity through indirect interconnections between entries of a document term matrix[30] and were able to create distributional thesauri with an (analog) computer. Borko & Bernick (1963) categorized documents based on an eigenvalue decomposition of the correlation between two document term matrices. They only used the document vectors, but not the term vectors—the latter would have been a form of word embeddings with documents serving as contexts. Switzer (1965) proposed representing terms with vectors containing their association with a limited set of terms. These were to be selected for having a high variation in their association with other terms. This is the earliest case of automatically derived low dimensional word[31] representations as an explicit goal that I am aware of. Switzer can thus be argued to be the inventor of word embeddings, but did not provide any implementation.

Another candidate for the invention of word embeddings is Koll (1979) who created a shared seven-dimensional vector space for documents and terms through an iterative process. Seven documents without any thematic overlap were used as anchors, each spanning one dimension. Words were then positioned based on their appearance in the initial documents, further documents based on the existing words and finally all words based on the now complete document collection.

Later, Deerwester et al. (1990) used SVD to reduce the dimensionality of a document term matrix—a method they called Latent Semantic Analysis (LSA)—and compared terms by calculating the cosine between their vector representations (Deerwester et al., 1990, pp. 398–399). Schütze also used SVD and experimented with multiple types of contexts, among those the now common options of a small window of adjacent words on both sides, but with a position-aware approach (Schütze, 1993; Schütze & Pedersen, 1993). He also identified all word vectors in U_d to be of equal importance (Schütze, 1992a). This was also discussed in multiple later studies (Bullinaria & Levy, 2012; Österlund et al., 2015) and lead to the insight that word vectors

[30] The document term matrix is closely related to the word context matrix and very popular in information retrieval. It describes connections between documents and terms assigned to them.

[31] Assuming 'word' and 'term' to be interchangeable, as (frequent) words occurring in a text are well-suited as index terms for it.

should not be weighted by their singular values in Σ_d (Levy et al., 2015). Schütze (1992b) also pioneered using sub-word derived embeddings by calculating the SVD between co-occurring character 4-grams and using those to represent words—a sub-word approach is also used in the recent `fastText`, a variant of `word2vec` with better support for rare words (Bojanowski et al., 2017; Mikolov et al., 2018).

Niwa & Nitta (1994) did not use SVD, but invented the PPMI word association metric and represented words with vectors of PPMI values between them and a set of middle frequency words. Lund et al. achieved memory efficient vector representations by discarding all co-occurring words except the 200 words with the highest variance for their co-occurrences with all investigated words (Lund & Burgess, 1996; Lund et al., 1995).

Current word embeddings are typically neural (Bojanowski et al., 2017; Mikolov et al., 2013a,b) or at least explicitly inspired by neural embeddings (Levy et al., 2015; Pennington et al., 2014).

The `word2vec` algorithms were developed by simplifying a recurrent neural network, i.e., one with a hidden layer aware of the state it was in while processing preceding input, for predicting a word in question based on its context words (Mikolov et al., 2013a,c). Earlier neural word embeddings were trained as parts of larger neural networks for performing some external task, e.g., predicting the likelihood of text segments or the part-of-speech of a word (see e.g., Bengio et al. (2003), Turian et al. (2010), Socher et al. (2011), Collobert et al. (2011), Al-Rfou et al. (2013)). More generally, they are based on research on learning embeddings for arbitrary objects and their relations with each other (Hinton, 1986).

Levy & Goldberg (2014c) showed that SGNS embeddings can be seen as an approximation of SVD applied to a PMI matrix, which lead to the transfer of several pre-processing and weighting options and the creation of SVD_{PPMI} (Levy et al., 2015).

Despite the widespread use of word embeddings, there are still many open questions about the structure of the resulting vector space. As mentioned before, Schütze (1992a) recognized that there are no most important dimensions and performance is robust even when (some) dimensions are removed. Mimno & Thompson (2017) showed SGNS word embeddings, but not GloVe ones, to be constrained to a small cone shaped portion of the embedding space and be oriented away from context word vectors.

The ability to compute analogies was tackled in several studies. Levy & Goldberg (2014b) and Gábor et al. (2017) explained it by reframing analogies as measuring pairwise word similarity, whereas Gittens et al. (2017) and Arora et al. (2016) provided explanations based on information theory. Arora et al. (2016) could thus show that the beneficial effect of dimensionality reduction—another unsolved question—is due to noise reduction, an explanation already suggested by Turney & Pantel (2010). They were also able to quantify this noise reduction, which is a first step towards the calculation of an optimal number of embedding dimensions—the number of dimensions is currently chosen based on rules of thumb, i.e., 200–300 dimensions for neural and 300–500 for SVD word embeddings (Levy et al., 2015; Mikolov et al., 2013a; Pennington et al., 2014). Patel & Bhattacharyya (2017) could show minimum dimensionality to be linked with the maximum number of words with the same similarity values before dimensionality reduction—dimensionality must be high enough to ensure these words can be equidistant in embedding space after dimensionality reduction.[32]

There are also efforts to improve embedding quality by adding similarity information from manually curated resources (such as WordNet; Miller (1995)) or combining multiple types of contexts (Faruqui et al., 2015; Park & Myaeng, 2017). Automatically creating similarity resources for enhancing word embeddings was suggested by Ferret (2017), but it remains unclear how this differs from an ensemble of different types of word embeddings (see e.g., Muromägi et al. (2017)). An interesting way to avoid the problem of choosing the right type of context(s) is performing dimensionality reduction on a tensor[33] containing different types of contexts at once. However, this approach is limited by computational complexity (Baroni & Lenci, 2010). Simply representing all types of context in different columns of a matrix is not equivalent, since information on interdependence is lost.

[32] While at least $n - 1$ dimensions are necessary for n equidistant points by euclidean distance, e.g., two for a triangle, this relationship is far more complex for cosine-similarity (Patel & Bhattacharyya, 2017, p. 33), making this approach unusable in practice.

[33] As a generalization of a matrix, a tensor of order n contains values addressed via an n-tuple, e.g., as T_{ijk} for a tensor T of order 3.

Another main research direction is the treatment of ambiguous words—should there be a specific embedding vector for each sense or can one vector model all senses. A single representation for all senses can be seen as problematic, since synonyms for all senses of an ambiguous word will become close in vector space, e.g., *pollen* and *refinery* become falsely similar due to their connection with the different senses of *plant* (Neelakantan et al., 2014, pp. 1059–1060). Training sense specific embeddings makes it necessary to automatically detect the correct sense of each word (see e.g., Reisinger & Mooney (2010), Neelakantan et al. (2014), Wang et al. (2015)). As in the case of more sophisticated contexts, benefits seem to be very task specific and small or nonexistent (Kober et al., 2017; Neelakantan et al., 2014). A major challenge during disambiguation is deciding on the correct number of senses for each word—most systems avoid this decision by using the same number of senses for all words (see Biemann (2006) for a counter example). So far, large performance improvements were only shown for systems with a linguistically implausible fixed number of senses (Lee & Chen, 2017; Neelakantan et al., 2014). In addition, very recent work by Dubossarsky (2018, pp. 58–67) showed the beneficial effect of sense specific embeddings to likely be an artifact. A similar increase in performance could be achieved by randomly assigning word senses, probably due to multiple embeddings working as some kind of ensemble.
Finally, very recently a new line of research started to question the reliability of word embedding algorithms and thus their suitability for qualitative research. See Chapter 4 for details.

2.4. Evaluating Similarity

The performance of word similarity measurements can be judged either intrinsically or extrinsically (see e.g., Schnabel et al. (2015)). Intrinsic evaluation investigates how well predictions match human judgments. In contrast, extrinsic evaluation tests how similarity judgments (or the vector space model these are derived from) used as features influence performance on some task. The predictive power of intrinsic evaluation on extrinsic task performance seems to be limited, making it good practice to perform both kinds of evaluation where possible (Batchkarov et al., 2016).

Human judgments for intrinsic evaluation are commonly collected through questionnaires and averaged between multiple judges to form re-usable data sets. Rubenstein & Goodenough (1965) were the first to collect such a set of 65 English noun pairs, which was later translated and re-annotated for German by Gurevych (2005), but is no longer relevant due to its small size. Finkelstein et al. (2002) developed WordSim-353, a still widely used test set with 353 English noun pairs.[34] While the data sets mentioned so far used manually selected pairs, Radinsky et al. (2011) used 287 pairs of English words with arbitrary parts-of-speech that co-occurred in the same newspaper articles to construct the MTurk[35] data set. MTurk was annotated through crowdsourcing, i.e., using an online platform to recruit paid annotators. Bruni et al. (2012) also used crowdsourcing to annotate 3,000 word pairs for the MEN[36] data set, pairs being sampled from a large collection of volunteer provided image labels. MEN differs from other data sets in judging word pairs against each other instead of directly asking judges for the similarity of a single pair, e.g., prompting them to decide if *parrot* and *pelican* are more or less similar than *automobile* and *car*. These data sets used relatedness oriented instructions (see page 17), however SimLex-999 by Hill et al. (2014) was annotated for strict similarity. It was created with crowdsourcing and covers 999 English word pairs with identical part-of-speech.
Gurevych et al. created several data sets[37] for German (Gurevych, 2005; Zesch & Gurevych, 2006). The largest thereof is Gur350 with 350 word pairs across different parts-of-speech, e.g., comparing *Afrika* 'Africa' with *historisch* 'historical'.
A less direct and psycholinguistically interesting way to gather word similarity judgments is measuring the influence of word pairs on reaction time in lexical priming studies, semantically related words being known to decrease reaction time (Lund & Burgess, 1996; Lund et al., 1995). A non-reusable example for direct collection of human judgments is the work by Schnabel et al. (2015), who used crowd-

[34] `http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/` [Accessed May 28th 2019].

[35] `http://www.kiraradinsky.com/files/Mtruk.csv` [Accessed May 28th 2019].

[36] `https://staff.fnwi.uva.nl/e.bruni/MEN` [Accessed May 28th 2019].

[37] `https://www.informatik.tu-darmstadt.de/ukp/research_6/data/semantic_relatedness/german_relatedness_datasets/index.en.jsp` [Accessed May 28th 2019].

sourcing to directly compare word similarity systems.
Human judgments contained in pre-existing resources can also be re-used, e.g., questions from language test material (see e.g., Landauer & Dumais (1997), Turney (2001)) or thesauri transformed to provide explicit similarity values (see e.g., Rada et al. (1989), Lin (1998), Pedersen et al. (2004)).
Word embeddings lead to an ongoing search for alternative intrinsic evaluation approaches. Analogy-based evaluation is commonly used and supported by two large data sets for English (see page 24). Konkol et al. (2017) recently proposed to evaluate word embedding algorithms by their suitability for predicting distances between geographical locations, whereas Gamallo (2018) suggests using an outlier detection task.

2.5. Diachronic Distributional Research

The collection of large digital corpora (see Chapter 3 for examples and specific pitfalls) has made diachronic corpus-based studies increasingly popular in linguistics (see e.g., Biber et al. (2000, p. 205)), computational linguistics and the digital humanities. This section discusses such approaches with a focus on recent studies using word embeddings. Using distributional methods to analyze diachronic processes was already suggested by Koll (1979, p. 48):

> "An author's development and changing interests could be plotted and traced. One could view the emergence of new disciplines or subdisciplines and the converging and diverging of older schools of thought. The changing focus of an organization could be followed. In short, the [semantic] space could serve as a visual history of the concept relations of its population."

Following de Saussure's classical signifier-signified model (Wunderli, 2013), two types of semantic change are possible, i.e., a fixed signifier (word) with a variable signified (meaning) or a fixed signified expressed through variable signifiers. An example for the former would be *mouse* getting the additional sense 'input device', whereas *Joseph Ratzinger* becoming *Pope Benedict XVI* is an example for the latter (from Tahmasebi et al. (2012)). It is further possible to distinguish between changes of the signified affecting primarily its denotation (e.g., the above-mentioned *mouse*) and those affecting primarily its emotional connotation (e.g., Pinker (1994)'s 'euphemism treadmill'). The latter was largely ignored so far and is discussed in Section 5.1.

Section 2.5.1 contains (mostly older) examples for studies not utilizing word embeddings, whereas Section 2.5.2 describes word embedding-based studies in general. Section 2.5.3 is concerned with studies trying to find general laws of semantic change based on word embeddings. Finally, Section 2.5.4 provides a short look at the assessment of validity in diachronic research. Applied research on tools for diachronic studies is discussed separately in Section 5.2.4.
Further discussion with a focus on technical advances in using word embeddings to study semantic change can be found in a very recent survey by Kutuzov et al. (2018) which was helpful for compiling this overview. Their survey is more optimistic than this thesis, as it focuses on work analyzing increasingly shorter time spans (see also Del Tredici et al. (2019)) and narrower domains, while largely ignoring validity and reliability issues. Especially the latter are, as shown in Chapter 4, severe for methods involving neural networks and should be cause for skepticism towards many results discussed here, in particular those discussing changes for single words.

2.5.1. Non-Embedding Approaches

Probably the simplest distributional approach is observing changes in word frequency. The premiere example for this line of research is Michel et al. (2011), who collected the Google Books N-Gram Corpus (see Chapter 3) and proposed using such large scale corpora to analyze all kinds of cultural processes, e.g., censorship or career trajectories. Frequency information is popular in corpus linguistics, e.g., Hilpert & Gries (2009) discussed clustering as a solution for dividing frequency time lines in segments. Frequency information is also used in the digital humanities, e.g., to study the assumed importance of researchers via the frequency of their names (O'Sullivan et al., 2017). Frequency combined with part-of-speech information can be enough to identify some cases of semantic word change, e.g., the rise of *apple* (corpus was lower cased) as a company name and proper noun in contrast to *apple* as a fruit and common noun (Kulkarni et al., 2015). Unsurprisingly, frequency-based approaches perform worse than more complex ones (Gulordava & Baroni, 2011; Kulkarni et al., 2015). However, Englhardt et al. (2019) found adding frequency information to benefit word embedding-based change detection.
Word association (see Section 2.2.1) is used for corpus linguistic studies (e.g., Biber et al. (2000, p. 205 & pp. 265–268) or Taavitsainen

(2015)), with some authors fully adapting vector representations (del Prado Martín & Brendel, 2016; Hilpert, 2007; Hilpert & Perek, 2015; Perek, 2014). These studies are often concerned with the rise of new syntactical constructions and thus employ syntactical information to determine contexts, e.g., filtering for part-of-speech patterns to analyze constructions like *V the hell out of NP* (Perek, 2014). However, non-word embedding vectors containing association values can also be used to study semantic change (see Gulordava & Baroni (2011), Zou et al. (2013)).

It is also possible to use clustering-based word sense disambiguation algorithms to track semantic change. Tahmasebi et al. linked clusters created for different time spans to find changes in word meaning (Tahmasebi et al., 2012; Tahmasebi & Risse, 2017a,b; Tahmasebi, 2013). A similar approach[38] was used in several other diachronic studies (Mitra et al., 2015, 2014; Riedl et al., 2014). They distinguished, in contrast to Tahmasebi et al., between different types of semantic change while comparing clusters over time. Recchia et al. (2017) clustered words based on pre-existing word embeddings (from Hamilton et al. (2016c)). Clusters were initiated based on embedding derived similarity during a first time span and then iteratively updated for each subsequent time span. Pölitz et al. (2015) used diachronic topic modeling on corpora filtered to contain only contexts of a word in question. The resulting topics were assumed to correspond to senses and used to track semantic change. The evaluation of sense specific word change is complicated due to differing levels of granularity. For example, Tahmasebi & Risse (2017b) had to manually merge their (too) fine-grained clusters, e.g., multiple clusters concerned with *rock* in the sense of music, for comparisons with a dictionary.

2.5.2. Embedding Approaches

The earliest example for embedding based diachronic research I am aware of is Sagi et al. (2009, 2012) tracking e.g., the semantic widening of *dog* during Early Modern English. They used SVD to generate time span specific embeddings based on a co-occurrence window, filtering words to retain only those of high–medium frequency. The resulting embeddings were then used to analyze whether the contexts of a

[38] They used Chinese Whisper clustering (Biemann, 2006), whereas Tahmasebi et al. used curvature clustering (Dorow, 2006).

word in question became more or less diverse—according to the average pairwise cosine between vectors representing context words—over time.

Another early example is Jurgens & Stevens (2009) using random indexing[39] on blog posts to generate word embeddings. They compared embeddings from subsequent time spans directly (by cosine) to judge whether a word had changed or not, a now extremely common approach. Kim et al. (2014) were the first to use `word2vec` embeddings in a diachronic setting and also developed a visualization with line charts of most similar words over time.[40] They, as well as several follow-up studies (e.g., Kulkarni et al. (2015) or Hamilton et al. (2016c)), used sub-corpora of the Google Books Ngram corpus (see Section 3.3) which are far larger than the corpora available beforehand.

Word embeddings need to be aligned, i.e., the same type of semantic information must be encoded in matching dimensions, to be comparable with each other. This is due to the embedding algorithm being free to use an arbitrary dimension to encode some semantic information—recall the information preserving rotation illustrated in Figure 2.3. Embeddings trained on different corpora, e.g., from different time spans, are not aligned without further intervention.[41]

Alignment can be achieved by using the embeddings from each time span to initiate those of the succeeding one (Kim et al., 2014), but such an approach greatly increases effective training time, as all time spans must be modeled sequentially.

Alternatively, word embeddings for each time span are trained independently and then aligned in a post-processing step (Hamilton et al., 2016c; Kulkarni et al., 2015; Szymanski, 2017; Zhang et al., 2015, 2016). Section 4.3 explores the influence of both approaches on word embedding reliability. Post-processing can either rotate all vectors at once to minimize the distance between corresponding embeddings (can be solved as an orthogonal procrustes problem by applying SVD, see Hamilton et al. (2016c); Schönemann (1966)) or modify each

[39] Random indexing (Kanerva et al., 2000) represents contexts with random vectors and words with vectors generated by calculating the sum/centroid of their contexts' vectors.

[40] Now used in JeSemE, see Section 5.2.4 for a discussion of visualizations.

[41] Due to the reliability issues discussed in Chapter 4, such an alignment can be necessary even when comparing embeddings trained on the same data.

vector separately. The latter approach uses regression models based on the assumption that each words' distance to reference words (e.g., its most similar words) should be temporally stable (Kulkarni et al., 2015; Szymanski, 2017; Zhang et al., 2015, 2016).

A recent solution to the alignment problem are embeddings trained simultaneously for multiple time spans. This is conceptually attractive, as two-step procedures can be argued to be an approximation of such a simultaneous process (Yao et al., 2018). Jatowt & Duh (2014) concatenated co-occurrence counts for multiple time spans in one matrix (i.e., using time span specific rows, respectively columns, to encode co-occurrences) and performed SVD to create time specific embeddings. Bamler & Mandt (2017) built upon SGNS which they extend with variants for sharing word and context vector information either only forward or both forward and backward in time, thus achieving smoother change trajectories than prior approaches. Rudolph & Blei (2018) extend a generalized form of CBOW to share word, but not context vectors, forward through time. Yao et al. (2018) used embeddings based on the decomposition of PPMI co-occurrence matrices. Their optimization process tries to balance two goals, i.e., embeddings for different time spans being similar with each other and embeddings for each time span being suited for accurate PPMI matrix reconstruction. Finally, Rosenfeld & Erk (2018)'s approach is an extension of SGNS which jointly trains a time independent embedding for each word and a vector representing all words at once for a point in time. These can then be combined to form time specific word embeddings.

Aligned embeddings make it possible to directly detect semantic change by comparing embeddings trained on different time spans, but for the same word (see e.g., Kim et al. (2014)). They can also be used to identify words that replaced each other, e.g., *iPod* and *Walkman* (Zhang et al., 2015). This can be done either by directly comparing embeddings (Szymanski, 2017) or by also using information on most similar words (Zhang et al., 2015, 2016).

Alignment is not needed in diachronic studies observing changes in the most similar words instead of directly comparing embeddings from different time spans. Such an approach was not only used in Chapter 5, but also in several digital humanities studies (e.g., Jo (2016), Kenter et al. (2015)). Rodda et al. (2016) generated time specific matrices with similarity values between all words and used the

correlation between those matrices to identify words which underwent semantic change (by their low correlation). Eger & Mehler (2016) created vectors representing each word with its similarity to all other words of the vocabulary at a point in time. They then identified words that likely underwent semantic change by comparing these vectors and sorting them by their relative differences between different points in time. Hamilton et al. (2016b) compared such an approach with measuring the cosine between aligned vectors, finding it to be more sensitive to changes affecting nouns. Kutuzov et al. (2017) also compared a cosine based analysis of aligned embeddings and an analysis where the similarity of words to anchor words at different points of time was observed, finding both to perform roughly similar as features for text classification.
Another type of study possible without cross temporal alignment was conducted by Garg et al. (2018), who measured the similarity of embeddings for occupations with those for words indicating gender or ethnicity. They could quantify stereotypes and misrepresentations via comparison with historical workforce composition.

2.5.3. Laws of Semantic Change

Several recent studies tried to find general 'laws of semantic change' by comparing embeddings trained for the same word, but on texts from different time spans.
Xu & Kemp (2015) tested linguistic hypotheses on the development of synonyms, finding synonymous words to stay similar to each other and develop in parallel. Eger & Mehler (2016) showed words to change with a constant corpus-specific speed. However, this speed was higher for shorter time spans, possibly implying some form of artifact. Dubossarsky et al. (2015, 2016) created clusters from `word2vec` embeddings and found prototypicality, i.e., closeness to the center of these clusters, to prevent semantic change. Dubossarsky et al. (2016) found parts of speech to have major effects on the average change between time spans, e.g., verbs change 50% faster than adjectives. They also found frequency and word change to be nearly un-correlated, i.e., frequent and infrequent words change at the same speed. In contrast, Hamilton et al. (2016c) found high frequency words to change slower than low frequency words and polysemy to speed up semantic change. However, Dubossarsky et al. (2017) showed that such studies are

prone to artifacts. They created control corpora—for which the appropriate amount of semantic change was known—by mixing texts from different time spans or subsampling texts from one time span. They found reports on correlations between apparent change[42] and word frequency as well as polysemy to be mostly due to noise, disproving results from Dubossarsky et al. (2015), Dubossarsky et al. (2016) and Hamilton et al. (2016c).
Note that this line of research was concerned with word populations and not individual words and is thus of limited relevance for this thesis and arguably also for the validity of its methods.

2.5.4. Validity

Evaluating and comparing approaches for modeling word change is challenging, both in general and due to a lack of resources (Kutuzov et al., 2018). With the exception of studies looking at short term changes (mainly in social media, but see also Gulordava & Baroni (2011)), the human lifespan limits the availability of annotators speaking both the historical and the contemporary variety of a language. While the former type of studies is increasingly popular (Kutuzov et al., 2018), it is questionable whether approaches are directly transferable to longer time spans.
The intrinsic test sets from Section 2.4 are of limited use to assess the quality of embeddings trained on non-recent texts, since they are artifacts of the time of their creation. For example the German Gur350 contains *Internetseite* 'website' and *Stoiber*, the name of a then prominent German politician. English test sets are also affected, e.g., *Arafat* and *Maradona* are part of WordSim-353, while MEN lists *ipod-n.*[43]
Four evaluation strategies directly using word embeddings and suited for long term trends were proposed so far, here ordered by their assumed potential for future research:

[42] Measured with SVD_{PPMI} or raw co-occurrence frequencies. PPMI vectors without dimensionality reduction were less affected. SGNS was reported to be similarly affected as SVD_{PPMI} in a footnote.

[43] All words in MEN are provided in lower case and with a suffix indicating their part-of-speech.

- Exemplary words, e.g., picked for their change according to the approach in question or from examples in prior publications, can be used to compare predictions and expectations (e.g., Kulkarni et al. (2015), Hamilton et al. (2016c)). These approaches tend to focus on a small number of high frequency words, limiting both insights into many potential errors and statistical significance.

- Language change can be simulated by manipulating the underlying corpora, e.g., by replacing words with each other (Kulkarni et al., 2015) or merging them to a single pseudo-word (Rosenfeld & Erk, 2018). This approach can be used to create large test sets with words of different frequencies and with arbitrary change patterns, but it might suffer from artifacts.

- Cross-temporal equivalents, e.g., succeeding office-bearers, can be compared for their similarity at different points in time (Yao et al., 2018; Zhang et al., 2015). Such test sets can potentially be created automatically from public records, e.g., Wikipedia entries.

- Creation of a gold standard based on text samples for a word in question from different points in time (Schlechtweg et al., 2017). Annotators (possibly linguistic experts) judge these for change phenomena, e.g., metaphorical usage.

Approaches can also be evaluated extrinsically, i.e., by using word representations as features for some downstream task. Mihalcea & Nastase (2012) suggested a task where text samples must be classified for their temporal provenance. This task was later adopted by the SemEval-2015 challenge (Popescu & Strapparava, 2015). Kutuzov et al. (2017) used changes in word representations for locations in news texts to train a classifier to recognize military conflicts. Jaidka et al. (2018) showed models for characterizing social media users to depend on recent word embeddings. Yoon et al. (2018) showed information retrieval to benefit from word embeddings trained on texts contemporary with the texts to retrieve.
Overall, research seems to focus on potential applications without first ensuring methods to be well-suited, e.g., by creating proper gold standards or conducting in-depth case studies. Artifacts become a potentially severe confounding factor which might mislead linguists

or scholars with a limited awareness for ongoing research in computational linguistics.

Due to its focus on word embedding reliability, this thesis provides only limited experiments on validity. Synchronic validity was assessed with word similarity and analogy test sets, showing our training setup to be well-suited (see Chapter 4 and especially page 57). Diachronic validity was assessed directly through qualitative case studies in Section 5.3 which indicate variants of SVD_{PPMI} to be apt for tracking the meaning of words over time.

Chapter 3

Diachronic Corpora

Sufficiently large corpora are a requirement for distributional studies in general[1] and this thesis in particular. Creating digital diachronic corpora is labor intensive, as older texts are seldom digitized, and rife with legal problems, as many 20th century texts are protected by copyright. To avoid such issues, this thesis relies on several pre-compiled corpora in both German and English which are described in this chapter.
The proper selection of texts is a major problem during corpus creation (see e.g., McEnery & Wilson (1996, pp. 64–66), Hunston (2002, pp. 14–16, pp. 28–30)), with Biber et al. (2000, 246–253) describing three strategies:

Proportional selection of texts according to text production, e.g., if 100 poems and 20 novels were written, one would sample in a 5:1 ratio to form a corpus. A downside of this strategy is the risk of missing rare phenomena.

Stratified selection of texts to represent each (important) variation, e.g., if 100 poems and 20 novels were written, one would sample from both equally to form a corpus. Proper text selection is more challenging than with other methods.

[1] Sometimes the results returned by internet search engines are used as substitutes (e.g., Turney (2001)), but this approach suffers from very opaque and ever-changing data, reducing both its validity and reliability.

Exhaustive selection of texts includes all texts from a source or author, e.g., if 10 poems and 8 novels were written by an author, one would combine all to form a corpus. This strategy is well-suited for domain-specific studies, but results are not representative for language at large.

Improper text selection and changes in corpus composition can mislead analyses as they are hard to distinguish from real distributional data (Koplenig, 2017). The case study on *Romantik* in Section 5.3.2 shows such an interference which was caused by a change in corpus composition described in Section 3.3.
Another confounding factor besides text selection is the quality of text digitization. Optical character recognition (OCR) has troubles processing old documents, e.g., a case study on historical newspapers found up to 1/3 of all words to be misrecognized (Tanner et al., 2009). A typical OCR error in historical texts is medial-s '$\int$' being confused with 'f' (Lin et al., 2012, p. 174). A study on OCR errors' effect on distributional analysis indicates robustness against low to moderate (about 20%) levels of OCR errors for topic modeling (Walker et al., 2010), which might also be the case for other methods. Higher quality digitization can be achieved with manual transcription, especially with double-keying, i.e., independent transcription by two individuals.
Finally, corpora differ in the level of text normalization, meta data and annotation they provide. Normalization is especially important for diachronic studies, e.g., German did not have a generally agreed-on orthographic standard in the 19$^{\text{th}}$ century due to political fragmentation (Schmidt, 2007, p. 172). Fortunately, all corpora provide normalization, with the exception of the Google Books Ngram corpus discussed in Section 3.3. While such normalization is unlikely to be flawless, it should provide more uniform input for distributional modeling.
Section 3.1 introduces the large and well-curated Corpus of Historical American English. Section 3.2 describes the Deutsches Textarchiv Kernkorpus ['German text archive core corpus'], a well-curated resource for German, especially for texts from the 19$^{\text{th}}$ century. Section 3.3 is concerned with the Google Books Ngram Corpus. It is the largest existing diachronic corpus and has sub-corpora for several languages and domains. Finally, Section 3.4 describes the relatively small and domain-specific Royal Society Corpus.

3.1. Corpus of Historical American English

The Corpus of Historical American English (COHA; Davies (2012)) is a stratified corpus spanning from the 1810s to the 2000s, with the only major changes in its composition being the inclusion of newspaper texts from the 1860s on and an increase in size between the 1810s and 1830s (see Figure 3.1). COHA is unique in being at the same time "quite large — 100 times larger than any other structured corpus. But it is also well balanced by genre and sub-genre in each decade, and it has been carefully lemmatized and tagged for part-of-speech." (Davies, 2012, p. 122). COHA can be freely queried online[2] or downloaded with a license.
COHA is annotated with automatically determined tokens, lemmata and part-of-speech tags. These were manually checked on the type level for the 100k most frequent words. It also provides limited meta data for each text sample, i.e., author, year and title.[3] COHA consists of texts from several preexisting collections, e.g., Project Gutenberg (see Davies (2012, p. 125) for details), often converted with OCR. Davies used a quality control scheme in which texts were compared with contemporary texts from the same genre to discard those likely to be affected by OCR errors.
A potential problem when using COHA is 10 words out of every 200 being replaced with placeholders due to fair-use limitations of the underlying texts.[4] Later experiments ignore these replaced words or use the online version for examples which seems to be unaffected and provides convenient mapping between text samples and sources.

3.2. Deutsches Textarchiv

The Deutsches Textarchiv Kernkorpus (DTA) ['German text archive core corpus'] is the result of an ongoing effort to create a digital full text corpus of printed German texts from the 15th to the 19th century (Geyken, 2013; Geyken & Gloning, 2015; Haaf, 2016; Haaf et al., 2015; Haaf & Thomas, 2016). It is based on manually transcribed (mostly

[2] `https://corpus.byu.edu/coha/` [Accessed May 28th 2019].
[3] See here for details on composition and text samples: `https://corpus.byu.edu/coha/files/cohaTexts.xls` [Accessed May 28th 2019].
[4] `https://www.corpusdata.org/limitations.asp` [Accessed May 28th 2019].

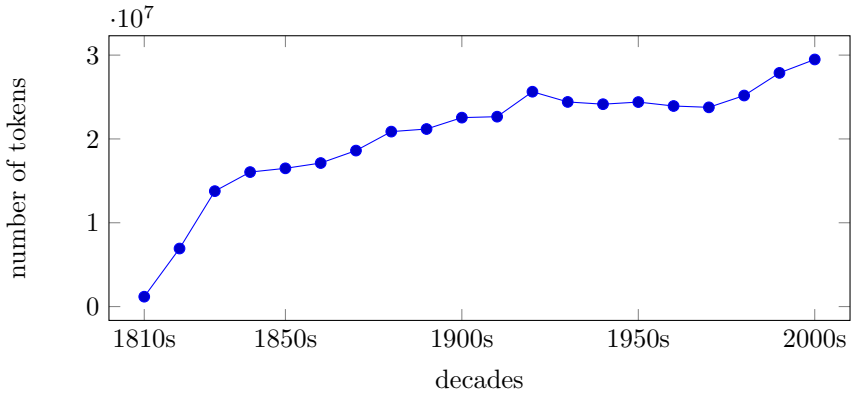

Figure 3.1: Number of tokens per decade in COHA.

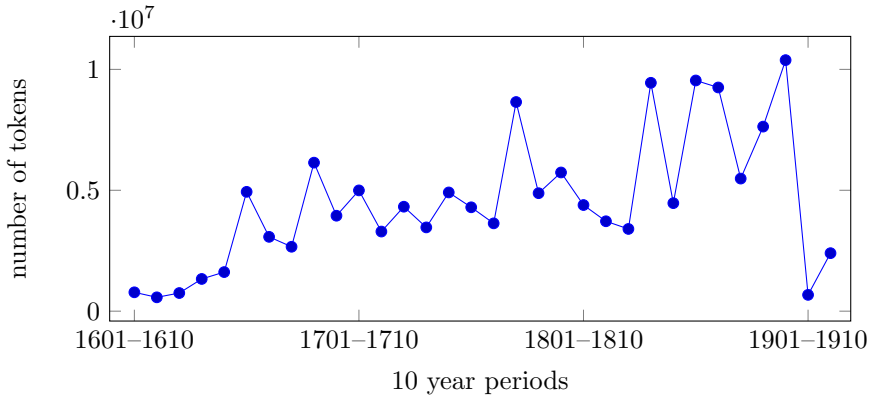

Figure 3.2: Number of tokens per decade in DTA.

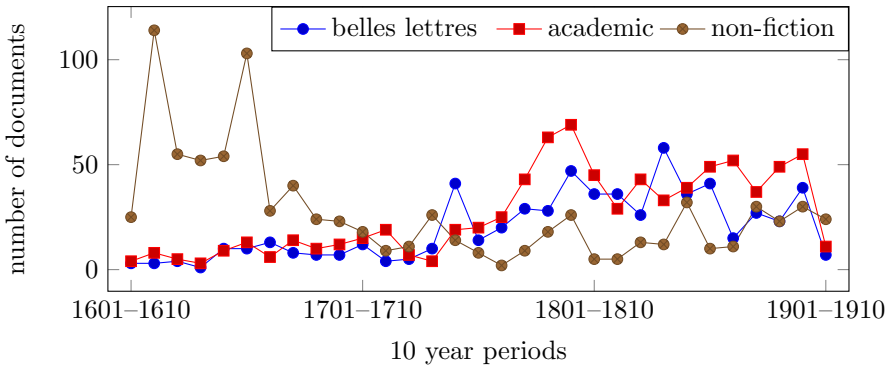

Figure 3.3: Composition of DTA over time.

double keying, in some cases corrected OCR) digital facsimiles and covers multiple genres, i.e., belles lettres, academic texts and general non-fiction (as well as a small number of journalism texts), and domains (e.g., poem, biology or medicine). It provides rich metadata on these categories as well as authors, publishers, editors, etc. and also offers automatic orthographic normalization (mapping archaic forms to contemporary ones) and lemmatization (Jurish, 2013). A snapshot of this still ongoing corpus project[5] is provided for download in the TCF format,[6] which preserves these metadata and linguistic annotations.[7] DTA is rather small and its size is not constant over time, as shown in Figure 3.2.
DTA aims to represent the overall development of German and can thus be described as a stratified corpus. Its individual texts were chosen for being representative from a linguistic point of view and can include historically popular translations.[8] The number of texts from each genre is not stable over time, as shown in Figure 3.3 based on online documentation. However, fiction and non-fiction are rather balanced for the timespans used in Sections 4 and 5.
The two main problems of using DTA for linguistic research are its only partially stratified nature and its relatively small size, which are addressed in later experiments by discarding early texts and combining texts from 30 years each (see Chapter 5).

3.3. Google Books Ngram Corpus

The Google Books Ngram corpus (Lin et al., 2012; Michel et al., 2011) contains texts in multiple languages and, for English, also multiple domains.[9] This thesis uses the English Fiction (GBF) and German

[5] The September 1st 2017 version was used here, see `http://www.deutschestextarchiv.de/download` [Accessed May 28th 2019].

[6] `https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format` [Accessed May 28th 2019].

[7] We developed tools for reading and converting this format and made them publicly available as part of the JULIE Lab UIMA Component Repository (Hahn et al., 2016; Hellrich et al., 2017): `https://github.com/JULIELab/dta-converter` [Accessed May 28th 2019].

[8] For example Lenz (1774) contains a translation of Shakespear's 'Love's Labour's Lost'.

[9] These are: English (with sub-corpora 'One Million', 'British', 'American' and 'Fiction'), Chinese, French, German, Hebrew, Italian, Russian and Spanish.

(GBG) sub-corpora. It is the largest diachronic corpus ever collected and contains about 6% of all books published between 1500 and 2009. Texts are provided as n-grams[10] up to 5-grams. Figure 3.4 shows the growth of GBF and GBG and the dominance of recent texts.

The Google Books Ngram corpus was created by digitizing library collections[11] as well as books provided by publishers. It is hard to describe with Biber et al. (2000)'s criteria (see page 49), as no information on the underlying books is provided. Research by Pechenick et al. (2015) showed its general language English sub-corpus to contain an increasing number scientific publications in recent years, with GBF being less affected. One indicator are parentheses which are more frequent in scientific texts due to their role in citations and references. Pechenick et al. (2015) did not analyze GBG, but the percentage of opening parentheses in Figure 3.5 clearly shows it to be highly affected—ignoring some very early outliers, there are far more parentheses in GBG than in GBF and their number increases over time. This combination of opaqueness and changing composition was already criticized in prior publications, as they could cause apparent changes in word usage (see e.g., Koplenig (2017)). The case studies in Section 5.3 confirm this warning, with GBG being very misleading, especially in regards to the word *Romantik*.

The Google Books Ngram corpus provides not only raw text, but also part-of-speech annotations and dependency parse tree fragments. Neither information was used in this thesis due to their limited utility when training word embeddings (see Section 2.1). The corpus can be searched for frequency information with a web interface[12] or be downloaded[13] for local processing, as was done here.

The n-gram format leads to sampling errors when contexts are determined with a sliding window (see Section 2.1), a fact barely discussed in the literature. If the outermost tokens of the n-grams are not

[10] Sequences of n tokens annotated with the frequency of this sequence in the corpus (per total occurrences and per books, the former being used in this thesis). For example a sequence of 5 tokens $a\,b\,c\,d\,e$ would result in the following 3-grams: $a\,b\,c$, $b\,c\,d$ and $c\,d\,e$.

[11] Based on a comparison of available titles, Jones (2010) found Google Books—of which the Google Books Ngram corpus is a subset—to be at least as well suited for 19$^{\text{th}}$ century studies as most American research libraries.

[12] `https://books.google.com/ngrams` [Accessed May 28$^{\text{th}}$ 2019].

[13] `http://storage.googleapis.com/books/ngrams/books/datasetsv2.html` [Accessed May 28$^{\text{th}}$ 2019].

used as centers, then their co-occurrences with inner words are undersampled, whereas the inverse is true if they are used as centers. This thesis uses the latter approach, while the former was used at least by Hamilton et al. (2016c).[14] Differences are likely to be small, as all words in the underlying texts are equally affected. A third alternative is only using the outermost words as centers, but none of the inner ones (Jatowt & Duh, 2014). This is problematic, as n-grams are calculated sentence-wise (Lin et al., 2012, p. 171), i.e., first and last words of a sentence are never used as contexts with this method.

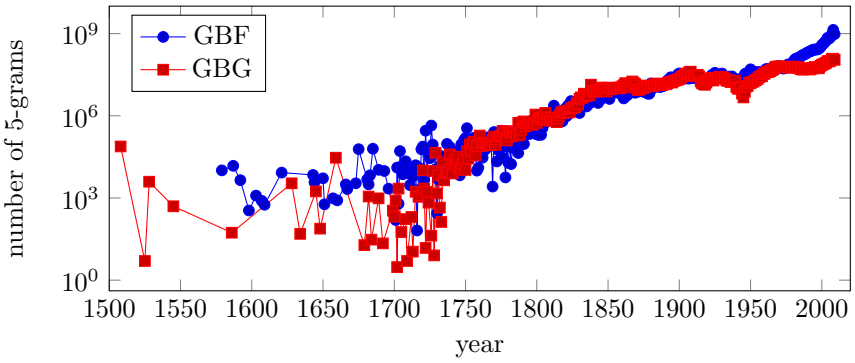

Figure 3.4: Number of 5-grams per year in GBF and GBG.

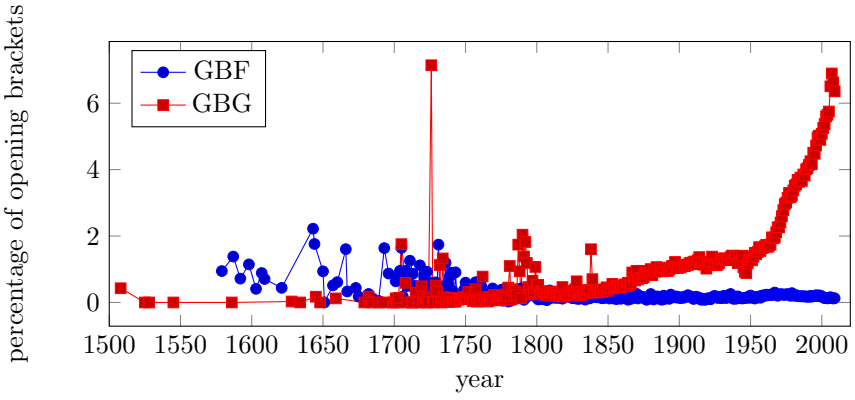

Figure 3.5: Percentage (per year) of 5-grams with at least one opening parentheses, i.e., '(', in GBF and GBG.

[14] Only evident from source code: `https://github.com/williamleif/histwords` [Accessed May 28th 2019].

3.4. Royal Society Corpus

The Royal Society Corpus (RSC; Kermes et al. (2016)) contains two centuries (1665–1869) of scientific publications from the Royal Society of London. It can be both queried online and downloaded[15] and includes the 'Philosophical Transactions' (i.e., the longest running scientific journal) and its spin-offs.[16]
Source documents were acquired from JSTOR[17] and were thus already scanned and processed with OCR. Kermes et al. (2016) used post-processing to correct OCR errors and performed automatic tokenization, lemmatization and part-of-speech tagging. They also added metadata (JSTOR already provided some, e.g., author and title) by using topic modeling to detect scientific disciplines and languages. RSC is pre-segmented into 50 year periods,[18] e.g., 1650–1699, with Figure 3.6 showing the respective size of each period.
RSC is thus an example for an exhaustive corpus and well suited for investigating questions in the history of science (see e.g., Section 5.3.1), but not for more general inquiries.

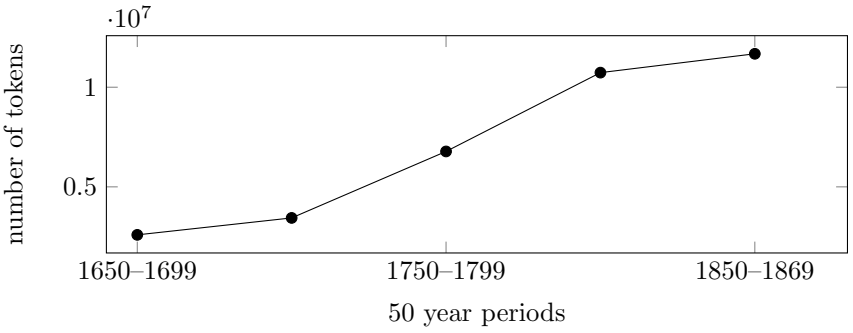

Figure 3.6: Number of tokens per time period in RSC.

[15] `https://fedora.clarin-d.uni-saarland.de/rsc/` [Accessed May 28th 2019].

[16] These spin-offs, i.e., 'Abstracts of Papers Printed in the Philosophical Transactions of the Royal Society of London', 'Abstracts of Papers Communicated to the Philosophical Transactions of the Royal Society of London' and 'Proceedings of the Royal Society of London', and a later split were required due to increasing submissions and the divergence of scientific disciplines (Fyfe et al., 2015, p. 233).

[17] `https://www.jstor.org` [Accessed May 28th 2019].

[18] The last one (1850–1869) is shorter, yet it contains more texts than others.

Chapter 4

Reliability of Word Embeddings

There are two fundamental requirements that word embeddings used for corpus linguistic research must fulfill. Firstly, they must be valid and model lexical semantics in a way that matches human understanding. Secondly, they must be reliable and produce consistent results when experiments are repeated.
In regards to validity, it is commonly measured intrinsically by solving analogies or judging word similarity with a word embedding model (see Section 2.4). We[1] used both strategies to gauge the suitability of the models trained in this chapter for down-stram tasks. In general, word embeddings with sufficient training material achieve about human level performance in word similarity tasks, unless evaluated for strict similarity (see Section 2.2.2). For example, the $\text{SVD}_{\text{wPPMI}}$ method trained on the 2000s time span of COHA achieved a rank correlation of $\rho = 0.35$ on SimLex-999 and $\rho = 0.58$ on WS-353 (see Section 4.6). The average pairwise correlation between annotators on both data sets is $\rho = 0.67$ and $\rho = 0.61$, respectively (Hill et al., 2014). With a larger training corpus—sadly seldom available for (diachronic) corpus linguistic studies—$\text{SVD}_{\text{wPPMI}}$ still performs below human level for SimLex-999 ($\rho = 0.44$), but above for WS-353 ($\rho = 0.65$).
In regards to reliability, we were, to the best of our knowledge, the first to investigate this problem. Section 4.1 describes how we quantified it in our experiments, while Section 4.2 provides background on causes

[1] Plural is used as experiments in this chapter were planed and published in co-operation with my supervisor Professor Dr. Udo Hahn and in one case also Bernd Kampe.

for (lacking) reliability. Section 4.3 contains our first reliability study on SG embeddings which was presented at the LaTeCH workshop (Hellrich & Hahn, 2016a). Section 4.4 describes follow-up research covering multilingual data and was presented at COLING (Hellrich & Hahn, 2016b). Section 4.5 provides an extension of our experiments to two additional algorithms, i.e., GloVe and SVD_{PPMI}. It allowed us to identify SVD_{PPMI} (with a proper downsampling strategy) as perfectly reliable and was presented at the Digital Humanities conference (Hellrich & Hahn, 2017a). The experiment in Section 4.6 is concerned with the influence of downsampling on reliability. We could show our novel SVD_{wPPMI} variant of the SVD_{PPMI} algorithm to be superior in reliability and comparable in accuracy. The statistical analysis for this experiment was performed by Bernd Kampe. A extended version of this experiment was presented at RepEval (Hellrich et al., 2019b). Finally, Section 4.7 provides an overview of the performance of all SGNS implementations used in this thesis and Section 4.8 discusses our results in relation to the scarce related work.

4.1. Quantifying Reliability

Reliability describes how well measurements can be repeated[2] and can thus be quantified by repeating a measurement and comparing results. Here, taking a measurement consists in training a word embedding model and using it to determine the most similar words for some words serving as anchors. A perfectly reliable word embedding method results in the same most similar words when trained multiple times on the same corpus (with the same parameters). In contrast, an unreliable method results in (some) different most similar words for these anchor words.
Several metrics can be applied to this problem which can be linked to older research on the agreement between distributional thesauri

[2] In a general sense, i.e., including both replication and reproduction. The former is concerned with carrying "out exactly the same task as the original researcher, with the expectation that the result will be the same" (Ivie & Thain, 2018, p. 63:3). In contrast, the latter consists of carrying "out tasks that are equivalent in substance to the original, but may differ in ways that are not expected to be significant to the final result" (Ivie & Thain, 2018, p. 63:4). The following experiments could be argued to be concerned with either replication or reproduction, depending on ones stance on seeds for random number generation being part of experimental setups.

(Padró et al., 2014; Weeds et al., 2004). Weeds et al. (2004) represented each anchor word with a vector containing its similarity (expressed as a rank) to all other anchor words. Systems were then compared by calculating the cosine between these vectors. Padró et al. (2014) compared the most similar words both with the Jaccard coefficient (Jaccard, 1912) as well as two position sensitive metrics.[3] Antoniak & Mimno (2018) also used the Jaccard coefficient, in addition to comparing the difference in rank for specific most similar words in relation to an anchor, as well as the variance in cosine between vector representations for an anchor and its most similar words. Wendlandt et al. (2018) used the percentage of identical words in the lists of most similar words for each anchor.

We used both a percentage based reliability metric (Section 4.3–4.5) as well as the Jaccard coefficient (Section 4.6). Both metrics are concerned only with words being among the most similar words for an anchor, but not their exact ranking. This is a deliberate choice due to the assumed corpus linguistic application, where scholars interpret a limited number of most similar words provided as lists or in visualizations (see Section 5.2 for example visualizations). The presence of a word among these (typically 2–10) most similar words is thus already a major factor for any interpretation, see Table 4.4 for an example of the differences to expect.

Our percentage based reliability metric for the n most similar words $r@n$ (Equation 4.1) and the averaged Jaccard coefficient for the n most similar words $j@n$ (Equation 4.2) are defined very similarly. Both depend on a set M of word embedding models for which the n most similar words (by cosine) for anchor words from a set A are compared, using the function $s(a, n, m)$ to provide the set of n most similar words for an anchor word a according to a model m:

$$r@n = \frac{1}{A} \sum_{a \in A} \frac{|\bigcap_{m \in M} s(a, n, m)|}{n} \tag{4.1}$$

$$j@n = \frac{1}{A} \sum_{a \in A} \frac{|\bigcap_{m \in M} s(a, n, m)|}{|\bigcup_{m \in M} s(a, n, m)|} \tag{4.2}$$

[3] Position sensitive metrics are not only concerned with the presence of a word among the most similar words, yet also with its relative position among those. An example is the cosine approach by Weeds et al. (2004).

The different divisor makes $j@n$ stricter than $r@n$, as non-identical words among the sets of most similar words cause both a lower dividend and a higher divisor. Increases in n seem to have relatively little effect on percentage based measures (and thus likely also $j@n$) for $n > 10$, with highest $r@n$ being measured for $n \approx 5$ (see below as well as Wendlandt et al. (2018)). Reliability values also differ with the size of M, more models leading to lower values (Antoniak & Mimno, 2018). Exact reliability values from different studies can thus not be compared in most cases, but general trends should be transferable.

4.2. Causes for Reliability Issues

Reliability issues can arise both from the creation of low dimensional word embeddings from high dimensional data[4] and from the preceding sampling of the high dimensional data. The latter case might seem relatively trivial—if contexts are downsampled (see Section 2.1) with a probabilistic approach, results will be different every time a program is run—, but its effect on embedding reliability was not investigated before the experiments described in Section 4.6. Problems due to the creation of low dimensional representations are probably more interesting from a theoretical point of view and lead to some surprising pitfalls in common implementations of word embedding algorithms.
These problems cannot affect SVD based word embeddings (unless stochastic SVD is used), as repeated SVD calculations produce identical singular vectors for a constant context matrix (Halko et al., 2011; Saad, 2003).[5]
In contrast, SGNS and GloVe both begin with random vectors which are then iteratively updated via stochastic gradient descent. Both, the initial starting position and the order in which examples are processed, can affect the final vectors. Stochastic gradient descent is well known to find one of many different local minima, but not a

[4] Either available explicitly in a word context matrix or implicitly in the examples processed by a streaming algorithm.

[5] Since SVD algorithms only approximate the exact results of analytic SVD, results are bound to a specific algorithm. The LAS2 algorithm used in all SVD word embedding experiments in this thesis is known for providing good singular vectors for high singular values, i.e., those kept during economic SVD (Berry, 1992). As with all computations different algorithm implementations and different computing hardware might also slightly affect outcomes, e.g., due to the limited precision of floating point numbers.

global one (see e.g., LeCun et al. (2015)). Thus, different staring positions are very likely to result in a different local minimum and thus different word embeddings every time the algorithm is run (recall the illustration in Figure 2.6).
It is important to remember that random numbers in computer programs are commonly not truly random, but only appear so despite being generated deterministically from an initial seed value (von Neumann, 1963). Using a fixed seed should make experiments deterministic and was suggested by e.g., Sandve et al. (2013) and Pierrejean & Tanguy (2018a). However, this would make the seed an additional parameter that can be optimized for performance, with at least some types of machine learning models being shown to appear significantly better or worse solely due to different seeds (Henderson et al., 2018). From a digital humanities point of view, the choice of seed (which determines the resulting most similar words) could lead to different qualitative interpretations.
Almost all implementations of word embedding algorithms evaluated in this thesis use such a fixed seed.[6] Their results are nevertheless nondeterministic due to the commonly used multi-threading confounding the order in which examples (in a corpus or word context matrix) are processed. Threads achieve different speeds and thus process different words when experiments are repeated.[7] This mismatch between algorithm and implementation can be assumed to be deliberate, as it eases debugging—a benefit of fixed seeds already highlighted by von Neumann (1963). Due to our results favoring SVD_{PPMI} (or its $\text{SVD}_{\text{wPPMI}}$ variant) we did not further explore the influence of the mismatch, except for the comparison of SGNS implementations in Section 4.7.

4.3. A First Look at SG Reliability

This first experiment is concerned with the accuracy and reliability of SG word embeddings derived from different training protocols as depending on word frequency, ambiguity and the number of itera-

[6] The only exceptions are the generation of initial vectors in GENSIM (only when PYTHON 3 is used) as well as downsampling in my modified version of HYPERWORDS.

[7] This can indirectly also causes different downsampling and potentially also different learning rates. See Section 4.7 for implementation specific details.

tions.[8] Its parameters and choice of corpus, i.e., samples of the GBF (see Chapter 3 for details), were inspired by two papers pioneering SG embeddings as a tool for diachronic analysis (Kim et al., 2014; Kulkarni et al., 2015).
A main difference between both papers is their way of solving the misalignment of vectors trained on different corpora, or in this case, diachronic sub-corpora for different years (see Section 2.5). Kim et al. (2014) trained models continuously, i.e., the word vectors for each time span (e.g., the year 1901) were initialized with the corresponding word vectors of the previous time span (e.g., the year 1900). In contrast, Kulkarni et al. (2015) trained models for each time span independently before aligning them.

4.3.1. Experimental Setup

Following the studies by Kim et al. (2014) and Kulkarni et al. (2015), we used a sub-corpus based on GBF and variations of their parameter choices. Kulkarni et al.'s protocol operates on all 5-grams in sub-corpora spanning five consecutive years (e.g., 1900–1904) and trains models independently of each other. In contrast, Kim et al. (2014)'s protocol trains on sub-corpora sampled to a uniform size of 10M 5-grams for each year from 1850 on in a continuous fashion, with years before 1900 used for initialization only. This constant size is achieved with a combination of under-sampling and over-sampling, as most years before 1880 do not achieve 10M entries (see also Figure 3.4 for corpus size per year). A comparison based on all years in the GBF would require several CPU years for training, thus we conducted an analysis on models for the beginning of the 20$^{\text{th}}$ century, i.e., 1900 for sample-based experiments and 1900–1904 for non-sampled ones (see Table 4.2 for size information). This choice was made since researchers can be assumed to be aware of current word meanings, making correct judgments on initial word semantics more important. We used the PYTHON-based GENSIM[9] SG implementation, which was easily modified for continuous training.[10] We trained[11] SGNS and

[8] Iterations over the entire training set, also known as 'epochs'.

[9] `https://radimrehurek.com/gensim/` [Accessed May 28$^{\text{th}}$ 2019].

[10] Code for experiments available from: `https://github.com/JULIELab/hellrich_latech2016` [Accessed May 28$^{\text{th}}$ 2019].

[11] 200 dimensions, symmetric 4 word context window, minimum frequency of 10 and 5 negative samples for SGNS.

SGHS[12] word embedding models. The threshold t for probabilistic downsampling of frequent words as well as the learning rate a were chosen in accordance with the studies by Kim et al. (2014) and Kulkarni et al. (2015) to be $t = 10^{-3}$ and $a = 0.01$ for sampled sub-corpora, respectively $t = 10^{-5}$ and $a = 0.025$ for the non-sampled sub-corpus. The learning rate is reduced during each iteration (down to 0.0001 for Gensim). Kim et al. (2014) did reset its value after each iteration, whereas Kulkarni et al. (2015) did not. We followed Kim et al. (2014)'s decision, as not resetting the learning reduces the impact of all iterations after the first.

Training was repeated to generate three models each for three protocols, i.e., training on the non-sampled 1900–1904 GBF, continuous training on sampled 1850–1900 GBF with all models trained on the same samples, and continuous training on sampled 1850–1900 GBF with all models trained on independent samples.

Training for each sub-corpus was repeated for 10 iterations, unless convergence was achieved before. Convergence was defined as the average cosine c between all corresponding word embeddings before and after a training iteration is 0.9999 or higher. The average cosine c_i for the ith iteration ($i > 1$) can be defined with a matrix W containing word embedding vectors (normalized to length 1) for words from a vocabulary of size V for each iteration i as:

$$c_i := \frac{1}{V} \sum_{v=1}^{V} W_{v,\,i} \cdot W_{v,\,i-1} \tag{4.3}$$

Accuracy was evaluated with the analogy test set developed at Google (Mikolov et al., 2013a), values equaling the percentage of correctly solved analogies. This test set is based on present-day English language and world knowledge, but it should allow for a comparison of models trained on older texts. Reliability was measured with the percentage based $r@n$ approach (see Section 4.1) with $1 \leq n \leq 5$. All words contained in any model for a corpus were used as anchor words. For anchor words not contained in a model, as can be the case for comparisons between different samples, $\emptyset$ was used as the set of most similar words.

[12] Kim et al. (2014) did not specify a SG variant, whereas Kulkarni et al. (2015) specified SGHS.

4.3.2. Results

Table 4.1 provides a general overview of the outcome of different training protocols. A first important observation is both accuracy and reliability being equal or higher for SGNS than for SGHS under all training protocols. In addition, reliability is higher between models trained on the same 10M sample than between models trained on the non-sampled corpus. However, if different samples are used—as during the comparison of independently reproduced experiments—reliability is dramatically lower for sample based approaches. Thus, a non-sampled approach (as used with SGHS by Kulkarni et al. (2015)) provides superior reliability in real word scenarios.
We also performed a more detailed analysis of models trained independently on a non-sampled 1900–1904 sub-corpus to quantify the influence of word frequency, ambiguity and the number of training iterations. Figure 4.1 shows the influence of word frequency, SGNS being overall more reliable (by $r@1$), especially for words with low or medium frequency. 21 exemplary words reported to have undergone

Training protocol			Reliability $r@n$					Accuracy
			1	2	3	4	5	
indep.	SGNS	non-sampled	0.40	0.41	0.41	0.40	0.40	0.38
		same sample	0.45	0.48	0.50	0.51	0.52	0.25
		different samples	0.09	0.10	0.10	0.10	0.10	0.26
	SGHS	non-sampled	0.33	0.34	0.34	0.34	0.34	0.28
		same sample	0.38	0.40	0.42	0.42	0.43	0.22
		different samples	0.09	0.09	0.10	0.10	0.10	0.22
cont.	SGNS	same sample	0.54	0.55	0.56	0.56	0.57	0.25
		different samples	0.21	0.21	0.22	0.22	0.22	0.25
	SGHS	same sample	0.31	0.32	0.32	0.32	0.33	0.22
		different samples	0.12	0.13	0.13	0.13	0.13	0.23

Table 4.1: Analogy accuracy and $r@n$ reliability for training with different parameters and algorithms for both **indep**endent and **cont**inuous training. Standard deviation for accuracy ± 0, except for independently trained SGHS where it is ± 0.01. Reliability is based on the evaluation of all lexical items, thus no standard deviation is given.

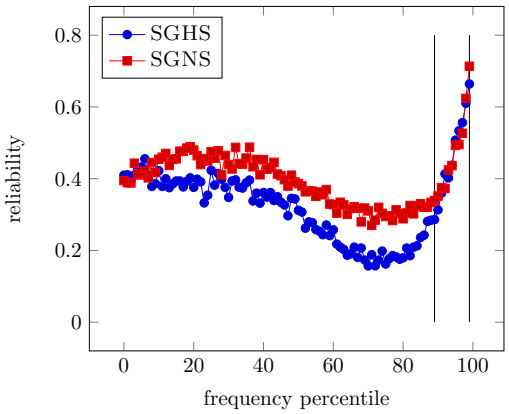

Figure 4.1: Influence of frequency percentile on r@1 reliability for models trained independently for 10 iterations on 1900–1904 GBF. Words previously reported to have changed during the 20$^{\text{th}}$ century fall into the rank range marked by vertical lines.

semantic changes in related work[13] belong to the highlighted 89–99 frequency percentile. Only for such high-frequency words does SGHS perform similar or slightly better. The dip in reliability for medium frequency words is further explored in Section 4.4.2.

Entries in a lexical database (here WordNet by Fellbaum (1998)) can be employed to measure the effect of word ambiguity on reliability.[14] The number of senses (i.e., WordNet synsets) seems to have little effect on r@1 reliability for SGNS, whereas the reliability of SGHS is lower for words with a low number of senses, as shown in Figure 4.2.

Both, reliability and accuracy, depend on the number of training iterations, as shown in Figure 4.3. There are diminishing returns for SGHS with reliability staying constant after 5 iterations, while SGNS increases in reliability with each iteration. However, both

[13] Kulkarni et al. (2015) compiled the following list based on prior work (Gulordava & Baroni, 2011; Jatowt & Duh, 2014; Kim et al., 2014; Wijaya & Yeniterzi, 2011): *card*, *sleep*, *parent*, *address*, *gay*, *mouse*, *king*, *checked*, *check*, *actually*, *supposed*, *guess*, *cell*, *headed*, *ass*, *mail*, *toilet*, *cock*, *bloody*, *nice* and *guy*.

[14] We used WordNet 3.0 and the API provided by the Natural Language Toolkit (Nltk): `http://www.nltk.org` [Accessed May 28$^{\text{th}}$ 2019].

methods achieve maximum accuracy after only 2 iterations, with continued training being harmful for SGNS performance.
Our overall result was a first warning to mistrust SG models as a novel corpus linguistic method. The training protocol by Kulkarni et al. (2015) appeared to be somewhat more reproducible than that of Kim et al. (2014). It was thus recommended by us to be used in further studies (Hellrich & Hahn, 2016a).

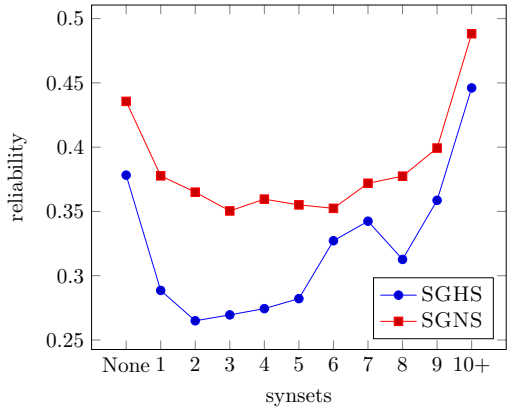

Figure 4.2: Influence of ambiguity (as number of WordNet synsets) on r@1 reliability for models trained independently with 10 iterations on 1900–1904 GBF.

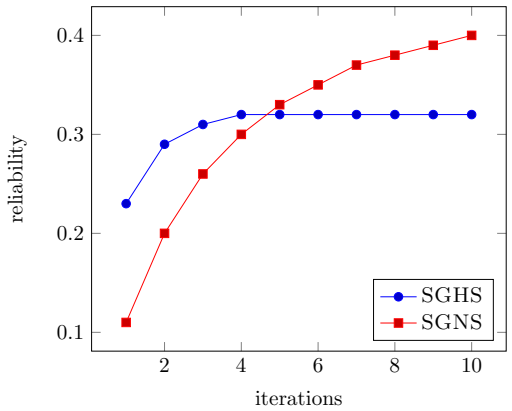

Figure 4.3: Influence of the number of iterations over training data on r@1 reliability, for 1900–1904 GBF.

4.4. A Multilingual View on SG Reliability

Follow-up experiments concern the accuracy and reliability of SG word embeddings trained not only on English texts, but also on German ones. We also widened our comparison to include both recent (2005–2009) and historical (1900–1904) texts. The general design of these experiments follows Kulkarni et al. (2015), i.e., embeddings are trained independently on non-sampled sub-corpora of GBF and GBG. Results confirm SG's fundamental lack of reliability and provided a first overview of the impact of normalization on word embeddings trained on historical German texts.

4.4.1. Experimental Setup

The training itself follows the approach described for the Kulkarni et al. (2015) protocol in Section 4.3.1, yet with an extended list of sub-corpora.[15] For both the 1900–1904 and the 2005–2009 period we trained three models each on GBF, GBG and a normalized GBG.
Normalization was achieved by generating a list of all types in GBG and applying CAB,[16] a tool developed for the normalization of the DTA (Jurish, 2013). The resulting mappings between types and the appropriate modern lemmata, e.g., archaic and inflected *medi<u>c</u>inische* to *medizinisch*, were then applied to the German sub-corpora in a pre-processing step. Normalization strongly reduces the number of types, as shown in Table 4.2. The typological contrast between both languages is especially clear for the 1900–1904 sub-corpora. They are very close in size, but without normalization German has 39% more types, whereas normalized German has 10% less than English. See also Figure 3.4 for general corpus size per year.
We also found the convergence criterion by Kulkarni et al. (2015) (see Equation 4.3) and our policy of resetting the learning rate after each iteration to interfere with each other, rendering the criterion ineffective. We thus introduced a new convergence measure Δc which we tracked to find a suitable threshold for potential follow-up exper-

[15] Code for experiments available from: `https://github.com/JULIELab/hellrich_coling2016` [Accessed May 28th 2019].

[16] `http://www.deutschestextarchiv.de/demo/cab/` [Accessed May 28th 2019].

iments. It is based on the change of c_i during subsequent iterations i ($i > 2$):

$$\Delta c_i := c_i - c_{i-1} \tag{4.4}$$

As in Section 4.3, we measured reliability with $r@n$ between three models, yet we also used word similarity based accuracy in addition to analogy based one. Similarity based accuracy was defined as Spearman's rank correlation coefficient between human similarity judgments (from a test set) and word similarity according to a SG model (cosine between word embeddings) between pairs of words. The main benefit of this approach is the existence of suitable test sets for both English and German in the form of WordSim-353 and Gur350 (see Section 2.4). Test pairs with words not modeled for a sub-corpus were ignored during evaluation.

Language	Time Span	5-grams	Types
English	1900–1904	143M	80k
English	2005–2009	4,658M	216k
German	1900–1904	135M	111k
Normalized German			72k
German	2005–2009	546M	243k
Normalized German			179k

Table 4.2: Number of 5-grams and types contained in the GBF and (normalized) GBG sub-corpora used here.

4.4.2. Results

Table 4.3 shows the overall similarity accuracy and $r@1$ reliability of our SG models, which allows for the following observations:

1. Both accuracy and reliability are higher for SGNS than for SGHS for all tested combinations of sub-corpora and time spans when trained for 10 iterations.

2. If only one iteration is used—as in many other experimental setups reported in the literature—there is little difference in accuracy while SGHS is clearly better in terms of reliability.

3. Accuracy is higher for 2005–2009 than for the 1900–1904 interval, with the exception of non-normalized GBG, probably due to temporal currency (see page 46) and larger size.

4. Normalization of German increases accuracy, but slightly decreases reliability.

Analogy accuracy could (for lack of test data) only be measured for GBF. Here we observed no negative effect of multiple iterations, but a more pronounced gap between algorithms. For example, 36% of all analogies were correct for SGNS and only 27% for SGHS after one iteration on the 1900–1904 sub-corpus, respectively 51% and 35% after one iteration on the 2005–2009 sub-corpus.
Like in Section 4.3, we also investigated the influence of several factors (e.g., frequency) in more detail, focusing on normalized GBG due to the overall similar performance and higher suitability for diachronic comparisons. We observed the same trends as before, e.g., the reliability for different numbers n of most similar words being very similar for a given combination of algorithm, sub-corpus and number of iterations. Figure 4.4 illustrates this for SGNS trained on the 1900–1904 GBF sub-corpus.

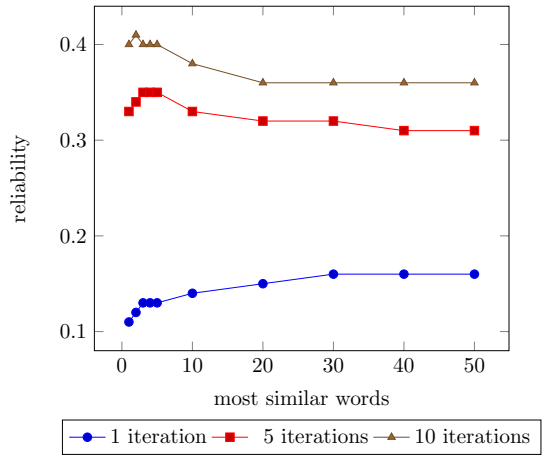

Figure 4.4: Effect of the number n of most similar words on $r@n$ reliability for SGNS trained on 1900–1904 GBF.

Training Scenario			r@1 Reliability			Similarity Accuracy		
Sub-corpus	Time Span	Algorithm	1 Iteration	5 Iterations	10 Iterations	1 Iteration	5 Iterations	10 Iterations
GBF	1900–1904	SGNS	0.11	0.33	0.40	0.45	0.51	0.51
		SGHS	0.23	0.33	0.33	0.46	0.45	0.45
	2005–2009	SGNS	0.36	0.54	0.57	0.58	0.58	0.57
		SGHS	0.36	0.39	0.38	0.55	0.52	0.52
GBG	1900–1904	SGNS	0.20	0.47	0.54	0.45	0.56	0.56
		SGHS	0.34	0.43	0.42	0.48	0.49	0.47
	2005–2009	SGNS	0.31	0.50	0.53	0.51	0.54	0.54
		SGHS	0.34	0.38	0.36	0.49	0.48	0.47
norm. GBG	1900–1904	SGNS	0.19	0.45	0.52	0.47	0.55	0.57
		SGHS	0.32	0.42	0.42	0.47	0.48	0.48
	2005–2009	SGNS	0.30	0.48	0.52	0.54	0.59	0.60
		SGHS	0.33	0.37	0.36	0.51	0.52	0.52

Table 4.3: Similarity accuracy and r@1 reliability for threefold repetition of different training scenarios after completing 1, 5 and 10 training iterations.

Word frequency was again linked with reliability, models being less reliable for medium frequency words than for high or low frequency ones. This effect was more pronounced for English than German, as shown in Figures 4.5 and 4.6. SGNS was more reliable than SGHS, especially for words with low or medium frequency.

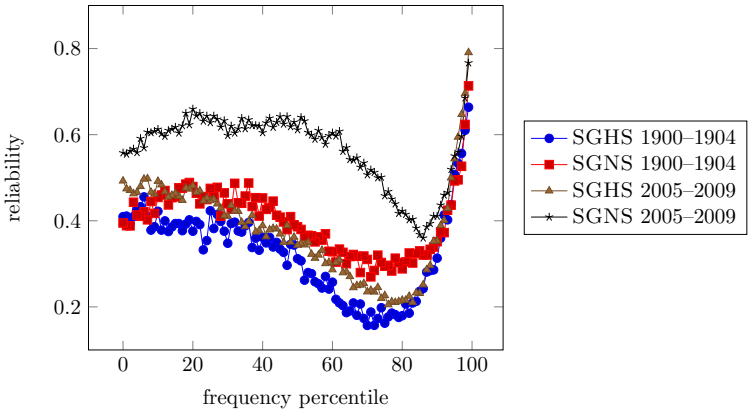

Figure 4.5: Influence of frequency percentile on reliability after training for 10 iterations on 1900–1904 and 2005–2009 GBF.

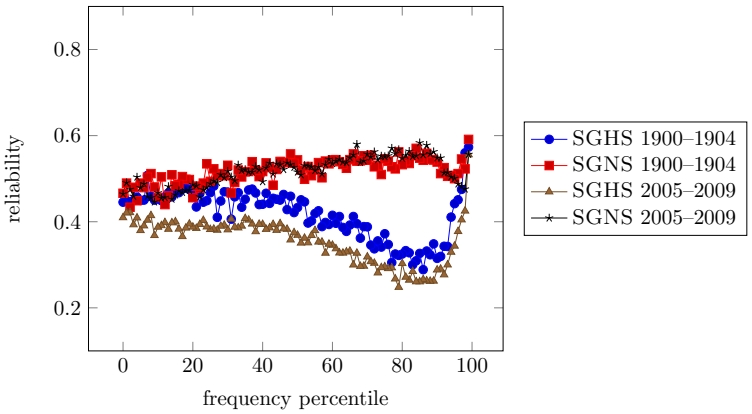

Figure 4.6: Influence of frequency percentile on reliability after training for 10 iterations on normalized 1900–1904 and 2005–2009 GBG.

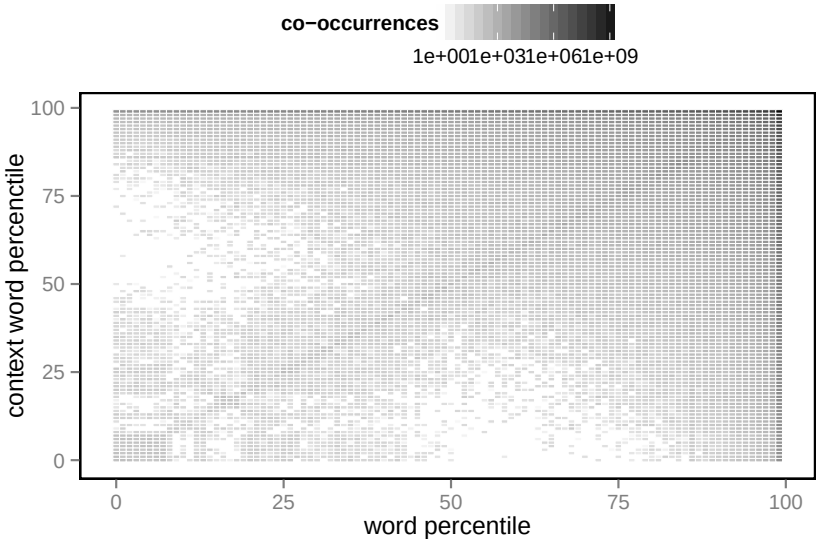

Figure 4.7: Number of co-occurrences (indicated by shade; only values above mode shown) between words and context words per frequency percentile for 1900–1904 GBF.

We assume the relatively low reliability for medium frequency English words to be caused by the pattern of word co-occurrences in the underlying corpus.[17] As shown in Figures 4.7 and 4.8, medium frequency words in GBF have fewer co-occurrences with low-frequency words than those in the normalized GBG. This might result in a lack of specific contexts for these words during training and thus reduce embedding quality.

Ambiguity was also again linked with $r@n$ reliability. In addition to WordNet (Fellbaum, 1998) we also used GermaNet[18] (Kunze & Lemnitzer, 2002) for German words. Highly ambiguous English words had better $r@n$, as shown in Figure 4.9. This effect was clearly

[17] Co-occurrences were counted with the hyperwords based ngrams2counts tool: `https://github.com/hellrich/hyperwords/blob/master/hyperwords/ngram2counts.py` [Accessed May 28th 2019].

[18] We used GermaNet 11.0 and the pygermanet API [Accessed May 28th 2019]: `https://pypi.python.org/pypi/pygermanet`

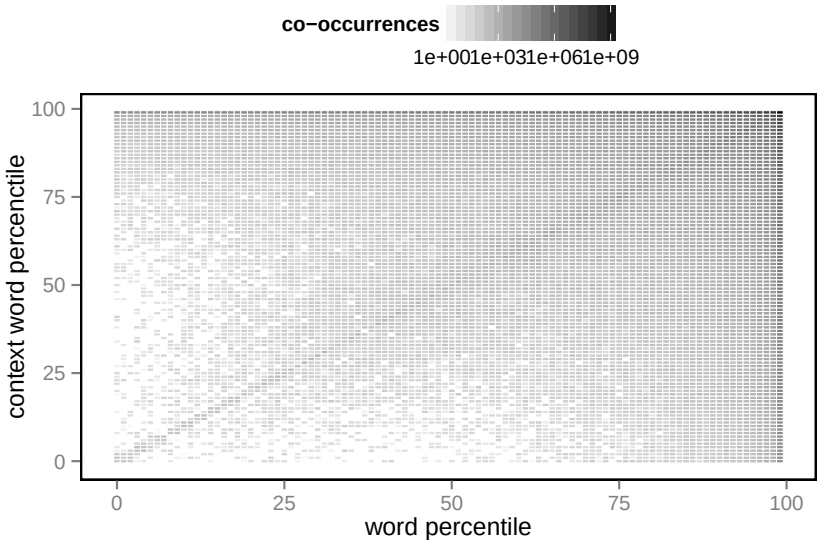

Figure 4.8: Number of co-occurrences (indicated by shade; only values above mode shown) between words and context words per frequency percentile for normalized 1900–1904 GBG.

reduced for German, as Figure 4.10 reveals. This counter-intuitive effect for English seems to be caused by the low ambiguity of infrequent words as results become more uniform when the analysis is limited to high frequency words—possibly due to typological differences.
Model reliability and accuracy depend again on the number of training iterations, as shown in Figures 4.11 and 4.12 for English, respectively normalized German. For both languages and time spans SGNS outperforms SGHS when training lasts for a sufficient number of iterations. The number of necessary iterations for SGNS to become superior seems to be linked to both language and corpus size, as it is lower for 2005–2009 than for 1900–1904 data. While reliability continues to increase for each subsequent iteration for SGNS, there are clear diminishing returns and even regression for SGHS.

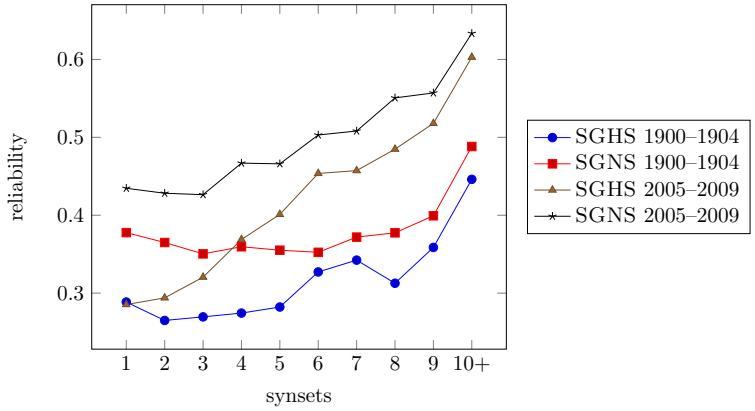

Figure 4.9: Influence of ambiguity (measured by the number of WordNet synsets) on r@1 reliability for models trained for 10 iterations on 1900–1904 and 2005–2009 GBF.

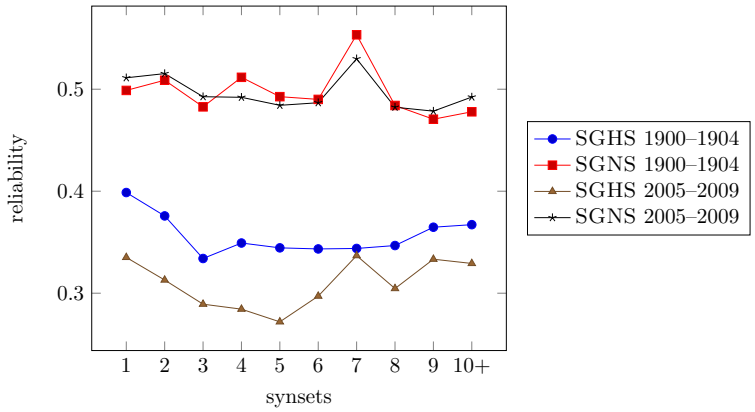

Figure 4.10: Influence of ambiguity (measured by the number of GermaNet synsets) on r@1 reliability for models trained for 10 iterations on normalized 1900–1904 and 2005–2009 GBG.

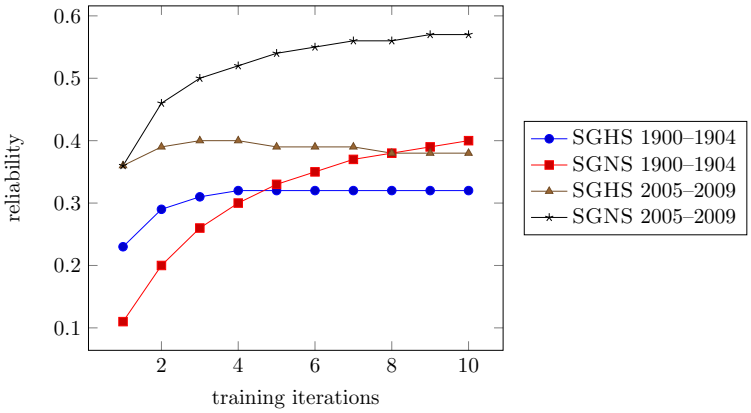

Figure 4.11: r@1 reliability as influenced by the number of training iterations for 1900–1904 and 2005–2009 GBF.

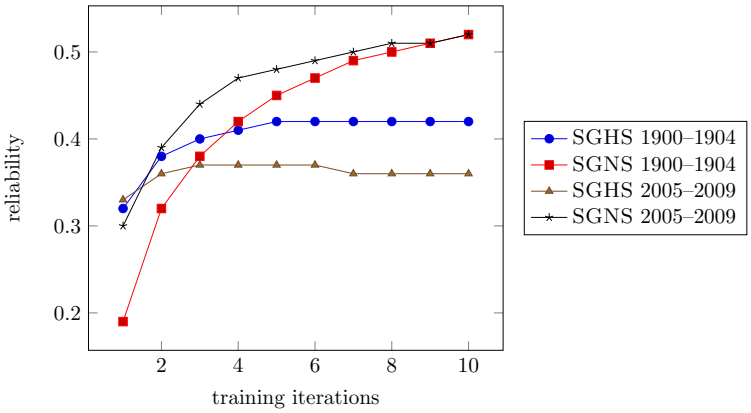

Figure 4.12: r@1 reliability as influenced by the number of training iterations for normalized 1900–1904 and 2005–2009 GBG.

To test for potential overfitting, we analyzed whether similarity accuracy was influenced by the number of iterations. Figures 4.13 and 4.14 show the results for English and normalized German, respectively. Note, that accuracy was assessed on a test set for modern-day language, limiting its validity. Accuracy also shows SGNS to benefit from multiple iterations, especially for smaller corpora. The biggest corpus (i.e., English Fiction 2005–2009) shows a slight regression in accuracy after more than 5 training iterations.
Overall both reliability and accuracy indicate SGNS with 4 to 6 iterations (6 being better for smaller and 4 being better for larger corpora) to be the best SG option for training word embeddings for our sub-corpora. Figure 4.15 shows Δc (see Equation 4.4) averaged over all three models between subsequent iterations, for both German and English SGNS models. Few changes occur after 4–6 iterations, which could be alternatively expressed as a Δc of about 0.003. We thus determined SGNS embeddings and convergence based on $\Delta c \lesssim 0.003$ to be the least bad choice.

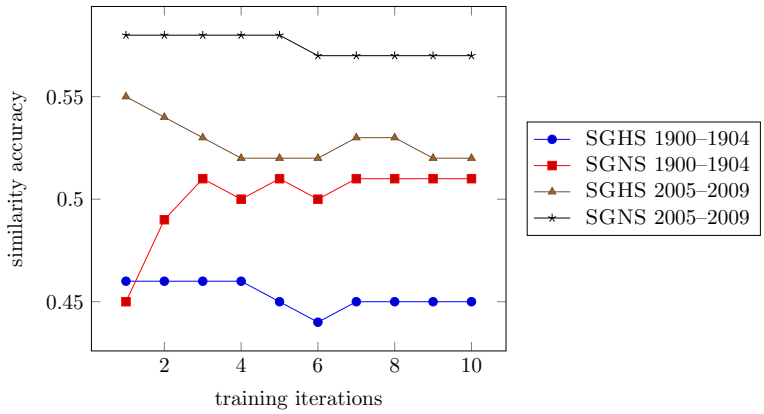

Figure 4.13: Similarity accuracy as influenced by the number of training iterations for 1900–1904 and 2005–2009 GBF. Error bars are not displayed on purpose, due to constant values for each training method.

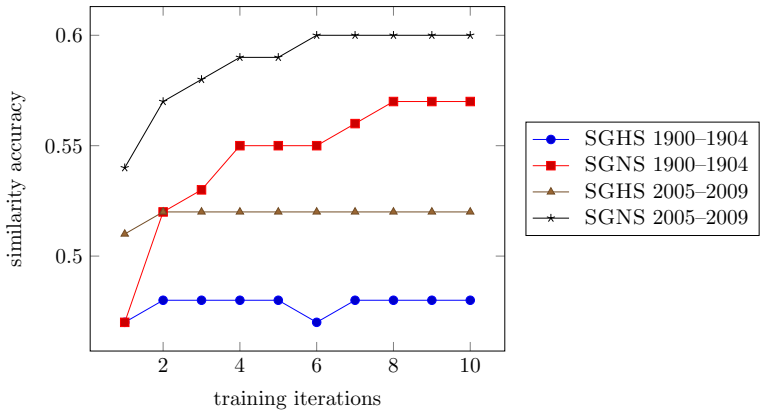

Figure 4.14: Similarity accuracy as influenced by the number of training iterations for normalized 1900–1904 and 2005–2009 GBG. Error constant for each training method, thus not shown.

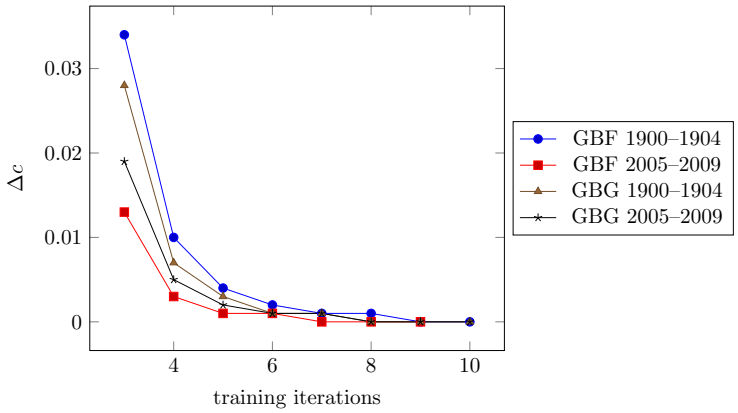

Figure 4.15: Change of averaged convergence Δc (between each iteration and its predecessor) for SGNS models trained on 1900–1904 and 2005–2009 normalized GBG or GBF.

4.5. Comparison of Word Embedding Algorithms

Prior experiments focused on the reliability of SG models only, due to their popularity in diachronic applications. In this experiment we widened our examination to also include GloVe and SVD_{PPMI}.

4.5.1. Experimental Setup

We used a data set consisting of all 645 German texts contained in the DTA for the 19th century (57M tokens). This allowed us to investigate reliability directly on a data set used for diachronic studies. DTA texts are provided normalized, further pre-processing consisted of casefolding and the removal of punctuation.
We trained three models each, now with three embedding algorithms, i.e., SGNS, GloVe and SVD_{PPMI}.[19] To increase reliability, we did only downsample word context combinations for GloVe, but not for SGNS and SVD_{PPMI}, as the latter two canonically use probabilistic downsampling—see Sections 2.3.2–2.3.4 for details on sampling in these algorithms and Section 4.6 for the effects of downsampling strategies on reliability.
Reliability was measured with $r@n$, accuracy was not collected as the rather old and thematically mixed provenance of the corpus makes test sets ill-suited. In contrast to previous experiments we used frequent nouns as anchor words instead of all modeled words. Reliability values are thus somewhat more optimistic and should match behavior for those high frequency words often studied in diachronic investigations.

4.5.2. Results

Overall, SVD_{PPMI} provided perfect reliability, while the other two embedding methods lacked in reliability. Table 4.4 shows the 1st to 5th most similar words for *Herz* as an example for the results of lacking reliability. The SGNS models diverge strongly, e.g., the first model finds *schmerzen* 'to pain' to be the most similar word, while the other two do not list it among the 5 most similar words. GloVe

[19] 300 dimensions, 5 word symmetric context window, minimum frequency threshold 5. 5 negative samples, one iteration (default) as well as 10 threads for SGNS. 8 Threads, 50 iterations (far more than for SGNS is common) for GloVe. Default learning rates.

appears to be more reliable, as all models agree on the first to third most similar words.
We also performed a quantitative analysis with different anchor words and various values of n for $r@n$. Figure 4.16 shows the reliability for each algorithm evaluated against the 1000 most frequent nouns in the DTA for $r@n$ with $1 \leq n \leq 10$. High values of n had a small positive effect on the reliability of SGNS and GloVe. A small inverse effect can be observed when the number of the most frequent nouns is modified while keeping a constant value of n, as displayed in Figure 4.17.
We could thus show GloVe to be affected by the same problems as SGNS and SVD_{PPMI} (without downsampling) to be perfectly reliable. Hence, we recommended it in Hellrich & Hahn (2017a) and used it for our own experiments in the next chapter.

Algorithm	Most Similar Word				
	1st	2nd	3rd	4th	5th
SGNS 1	*schmerzen* [to pain]	*beklommen* [anxious]	*busen* [bosom]	*bluten* [to bleed]	*herzen* [to caress]
SGNS 2	*bluten* [to bleed]	*klopfend* [beating]	*busen* [bosom]	*beklommen* [anxious]	*herzen* [to caress]
SGNS 3	*herzen* [to caress]	*busen* [bosom]	*klopfend* [beating]	*beklommen* [anxious]	*bluten* [to bleed]
GloVe 1	***gemüt*** [mind]	***mein*** [my]	***seele*** [soul]	*liebe* [love]	*brust* [chest]
GloVe 2	***gemüt*** [mind]	***mein*** [my]	***seele*** [soul]	*brust* [chest]	*liebe* [love]
GloVe 3	***gemüt*** [mind]	***mein*** [my]	***seele*** [soul]	*brust* [chest]	*liebe* [love]
SVD_{PPMI} 1	***busen*** [bosom]	***fühlen*** [to feel]	***liebe*** [love]	***schmerzen*** [pain]	***menschenherz*** [human heart]
SVD_{PPMI} 2	***busen*** [bosom]	***fühlen*** [to feel]	***liebe*** [love]	***schmerzen*** [pain]	***menschenherz*** [human heart]
SVD_{PPMI} 3	***busen*** [bosom]	***fühlen*** [to feel]	***liebe*** [love]	***schmerzen*** [pain]	***menschenherz*** [human heart]

Table 4.4: Most similar words for *Herz* [heart] as provided by different word embedding models. Words which all three models for an algorithm determined to be the nth most similar one in **bold**.

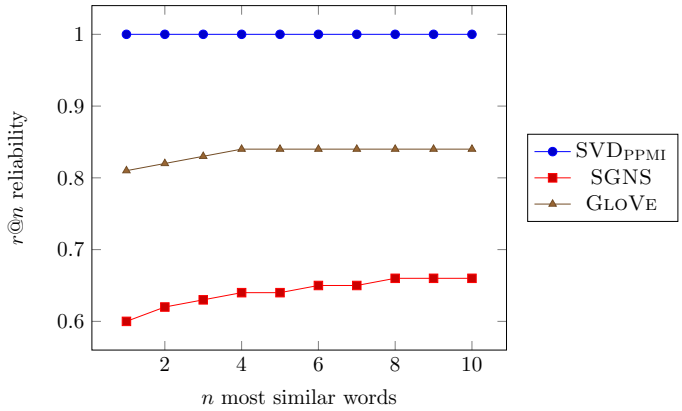

Figure 4.16: $r@n$ reliability of word embedding algorithms for $1 \leq n \leq 10$ and the 1000 most frequent nouns.

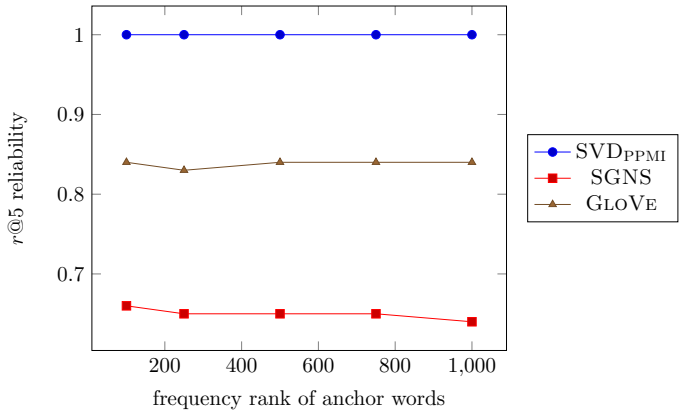

Figure 4.17: $r@5$ reliability of word embedding algorithms for the 100 to 1000 most frequent nouns used as anchors.

4.6. Downsampling and Reliability

Choosing between no downsampling, probabilistic downsampling and weight-based downsampling (see Sections 2.3.2–2.3.4 as well as 2.1) has a surprisingly large influence on the reliability of word embedding algorithms. Whereas Antoniak & Mimno (2018) used probabilistic downsampling for high frequency words and found SVD_{PPMI} to be less reliable than alternatives, we conducted the experiments in the preceding section without downsampling and found it to be perfectly reliable.

We compared the influence of different downsampling strategies on word embedding reliability and accuracy. Our tests spanned three algorithms (SVD_{PPMI} as well as GloVe and SGNS) and two English corpora of different sizes. We used these corpora both unmodified as well as bootstrap subsampled to simulate the arbitrary content selection in most corpora—texts could be removed or replaced with similar ones without changing the overall nature of a corpus. Bootstrap subsampling thus measures how trustworthy results are based on some corpus and which problems arise even with a perfectly reliable embedding method.

We added support for weighting-based downsampling to SGNS and SVD_{PPMI}. The latter approach, i.e., our novel $\text{SVD}_{\text{wPPMI}}$(see also page 28), outperforms prior SVD_{PPMI} variants. It uses fractional co-occurrence counts (according to sampling factors) to populate the word-context matrix before PPMI values are calculated and SVD is applied. Our weighted SGNS variant processes each word context pair (instead of discarding some as in probabilistic downsampling) and lowers the learning rate according to the appropriate sampling factors for the pair.

4.6.1. Experimental Setup

We used two different corpora for our analysis, i.e., the 2000s decade of COHA and an English News Crawl Corpus (NEWS) collected for the 2018 WMT Shared Task.[20] The former contains 14k texts and 28M tokens, while the latter has 27M texts and 550M tokens. Results on COHA are directly applicable for diachronic research, whereas the

[20] `http://www.statmt.org/wmt18/translation-task.html` [Accessed May 28th 2019].

larger NEWS corpus serves to gauge the performance of all algorithms in general applications.

Both corpora were tokenized, transformed to lower case and also stripped from punctuation. Bootstrap subsampling was performed on the level of the constituent texts of each corpus, e.g., individual news articles. For a corpus with n texts we drew n samples with replacement. Texts could thus be drawn multiple times, but only one copy was kept, roughly reducing corpora to $1 - 1/e \approx 2/3$ of their original size.

We trained ten[21] models with all algorithm variants each on the original corpora as well as on the independently bootstrap subsampled corpora. We measured similarity accuracy with Spearman's rank correlation between cosine and human word similarity judgments in MEN, MTurk, SimLex-999 and WordSim-353 (see Section 2.4). We also measured analogy accuracy as the percentage of correctly solved analogies (using the state-of-the-art multiplicative Equation 2.11) from two test sets developed at Google and Microsoft Research (Mikolov et al., 2013a,c).

Following Antoniak & Mimno (2018) and aiming to establish a common standard, we switch our reliability evaluation from $r@n$ to $j@n$—this should have little effect on overall rankings of algorithms as explained in Section 4.1. Reliability was measured as the $j@10$ Jaccard coefficient with the 1k most frequent words in each corpus (before bootstrap subsampling) as anchor words. We did not calculate $j@10$ between all 10 models for each combination of corpus, algorithm and downsampling strategy. Instead, we calculated $j@10$ between subsets of 9 models. 10 such unique subsets exist, allowing us to perform significance tests. The experimental code is available via GitHub.[22]

4.6.2. Results

Results for all tested combinations of corpora, algorithms and downsampling strategies are provided in Tables 4.5–4.8. Accuracy for smaller corpora (which includes bootstrapped ones) is lower than for

[21] 500 dimensions, symmetric 5 word context windows, minimum frequency threshold 50 for COHA, 100 for NEWS. Frequent word downsampling thresholds of $c = 100$ and $t = 10^{-4}$ where applicable. 5 negative samples for SGNS. Default learning rate, number of threads and number of iterations for each algorithm.

[22] `https://github.com/hellrich/embedding_downsampling_comparison` [Accessed May 28th 2019].

larger ones.[23] Also expected is the lower reliability for bootstrapped corpora, which is due to the difference in training material.

SGNS did seldom benefit from our modifications. Avoiding downsampling improved analogy (but not similarity) accuracy on COHA, while it was slightly worse on the larger NEWS corpus. Weight-based downsampling decreased accuracy, however it did benefit reliability, even when compared to training without downsampling.

$\text{SVD}_{\text{wPPMI}}$ provided perfect reliability in all non-bootstrapped scenarios. While its reliability in bootstrapped scenarios is slightly behind GloVe on COHA (0.329 instead of 0.33; difference significant with $p < 0.05$ by two sided t-test), its accuracy is higher. Notably, its reliability on the bootstrapped NEWS corpus (0.635) is nearly as high as that of probabilistic SGNS (0.652) and GloVe (0.679) trained on the non-bootstrapped corpus!

The overall accuracy of $\text{SVD}_{\text{wPPMI}}$ is about equal to the widely used probabilistic SGNS algorithm, i.e., it achieves higher values in 11 out of 24 measurements and a draw in 1. Its accuracy is higher than that of GloVe in 19 measurements and identical in 1. $\text{SVD}_{\text{wPPMI}}$ seems to benefit slightly from smaller corpus sizes (as for COHA), especially when compared with GloVe. This matches observations on embedding algorithms and corpus size made by Sahlgren & Lenci (2016).

The only prior option for perfectly reliable SVD_{PPMI} embeddings, i.e., using no downsampling at all as in Hellrich & Hahn (2017a), is beaten by $\text{SVD}_{\text{wPPMI}}$ in 20 out of 24 accuracy measurements with a draw in 3 out of 24. We could further confirm these results (on accuracy) by using the Friedman test (confirming the general existence of differences with $p < 10^{-6}$) followed by pairwise Wilcoxon rank-sum test with Holm-Šidák correction (see Demšar (2006)). We also found $\text{SVD}_{\text{wPPMI}}$ to be significantly better than SVD_{PPMI} without any downsampling as well as GloVe (for both $p < 10^{-3}$), but not better than SGNS ($p = 0.48$).[24]

[23] Standard deviations for accuracy are overall small (most around 0.01), with both bootstrapping and probabilistic sampling leading to increases (up to 0.03), and thus not listed in the tables. However, our comparisons are based on statistical tests and do thus use this information.

[24] Due to the correction, $\alpha = 0.0125$ was used (instead of the common 0.05) to determine significance. Bernd Kampe performed this part of the analysis.

Overall, using weight-based downsampling seems to be not worthwhile for SGNS, but well-suited for SVD_{PPMI}. The sometimes positive effect of probabilistic downsampling might be due to noise in neural networks working akin to dropout, i.e., it prevents overfitting and increases robustness (Goodfellow et al., 2016, p. 237). We assume that the overall low performance of weight-based downsampling for SGNS is not caused by vector updates becoming to small, as changing the learning rate could not compensate it during a limited pre-test. We deem $\text{SVD}_{\text{wPPMI}}$ to be the best embedding method for corpus linguistic (diachronic) research, since it performs as well as or better than other SVD_{PPMI} variants or SGNS (and far better than GloVe) while providing superior reliability.

Algorithm	Downsampling		Word Similarity				Analogy		Reliability
	Frequency	Window	MEN	MTurk	SimLex	WS-353	Google	MSR	
SVD_{PPMI}	none	prob.	0.681	0.567	0.322	0.581	0.237	0.253	0.440
		none	0.697	**0.582**	0.318	0.591	0.248	0.226	**1.000**
		weight	0.698	0.553	0.341	**0.596**	0.268	0.267	**1.000**
	prob.	prob.	0.689	**0.571**	0.333	0.577	0.224	0.257	0.324
		none	**0.703**	0.571	0.328	0.591	0.246	0.243	0.441
		weight	**0.703**	0.568	0.349	**0.593**	0.259	0.277	0.475
	weight	prob.	0.688	0.567	0.331	0.582	0.232	0.260	0.424
		none	0.702	0.560	0.328	0.589	0.247	0.243	**1.000**
		weight	*0.702*	*0.551*	*0.351*	*0.594*	*0.262*	*0.277*	**1.000**
SGNS	none	prob.	0.620	0.553	0.392	0.511	0.275	0.355	0.291
		none	0.645	0.567	0.355	0.535	**0.289**	**0.360**	0.305
		weight	0.619	0.541	0.375	0.512	0.185	0.228	0.201
	prob.	prob.	0.642	0.560	0.394	0.551	0.248	0.311	0.288
		none	0.660	**0.578**	0.347	0.561	0.259	0.301	0.282
		weight	0.620	0.532	0.372	0.516	0.162	0.144	0.234
	weight	prob.	0.579	0.496	0.387	0.487	0.255	0.323	0.567
		none	0.607	0.513	**0.407**	0.515	0.248	0.306	0.496
		weight	0.588	0.519	0.352	0.496	0.191	0.233	0.554
GloVe	weight	weight	0.590	0.522	0.222	0.405	0.167	0.214	0.808

Table 4.5: Performance of different algorithms and downsampling strategies with models trained on COHA without bootstrap subsampling. **Bold** values are best or not signtificantly different by t-test (with $p < 0.05$). Results for SVD_{wPPMI} in *italics*.

Algorithm	Downsampling		Word Similarity				Analogy		Reliability
	Frequency	Window	MEN	MTurk	SimLex	WS-353	Google	MSR	
SVD_{PPMI}	none	prob.	0.629	0.527	0.271	0.551	0.187	0.214	0.291
		none	0.645	**0.537**	0.267	**0.569**	0.192	0.184	0.310
		weight	0.645	**0.546**	0.295	**0.564**	**0.212**	0.228	0.325
	prob.	prob.	0.632	0.519	0.287	0.542	0.169	0.203	0.198
		none	**0.648**	**0.529**	0.280	**0.559**	0.185	0.188	0.217
		weight	**0.652**	**0.529**	0.307	**0.564**	0.204	0.228	0.251
	weight	prob.	0.635	**0.533**	0.284	**0.549**	0.180	0.212	0.274
		none	**0.649**	**0.536**	0.282	**0.559**	0.190	0.196	0.300
		weight	***0.651***	***0.534***	*0.305*	***0.568***	*0.206*	*0.235*	*0.329*
SGNS	none	prob.	0.535	0.470	**0.358**	0.456	**0.215**	**0.286**	0.123
		none	0.569	0.515	0.325	0.493	**0.219**	**0.285**	0.130
		weight	0.528	0.456	0.343	0.450	0.127	0.159	0.059
	prob.	prob.	0.551	0.486	**0.363**	0.479	0.192	0.243	0.091
		none	0.576	0.505	0.320	0.518	0.182	0.215	0.087
		weight	0.531	0.470	0.322	0.466	0.116	0.104	0.052
	weight	prob.	0.497	0.453	0.317	0.444	0.191	0.267	0.172
		none	0.529	0.479	0.352	0.466	0.195	0.260	0.139
		weight	0.508	0.461	0.278	0.440	0.122	0.171	0.131
GloVe	weight	weight	0.518	0.470	0.182	0.383	0.120	0.165	**0.330**

Table 4.6: Performance of different algorithms and downsampling strategies with models trained on bootstrap subsampled COHA. **Bold** values are best or not significantly different by t-test (with $p < 0.05$). Results for SVD_{wPPMI} in *italics*.

Algorithm	Downsampling		Word Similarity				Analogy		Reliability
	Frequency	Window	MEN	MTurk	SimLex	WS-353	Google	MSR	
SVD_{PPMI}	none	prob.	0.776	0.560	0.417	0.643	0.467	0.392	0.654
		none	0.775	0.559	0.406	0.643	0.469	0.357	**1.000**
		weight	0.776	0.561	0.417	0.642	0.473	0.395	**1.000**
	prob.	prob.	0.784	0.561	0.431	0.666	0.492	0.445	0.654
		none	0.786	0.572	0.423	0.668	0.504	0.420	0.801
		weight	**0.786**	0.569	0.434	0.668	0.504	0.449	0.806
	weight	prob.	0.783	0.561	0.433	0.666	0.492	0.440	0.681
		none	0.785	0.574	0.424	0.666	0.503	0.413	**1.000**
		weight	*0.786*	*0.568*	*0.435*	*0.667*	*0.502*	*0.444*	***1.000***
SGNS	none	prob.	0.726	0.648	0.429	0.641	0.630	0.560	0.631
		none	0.731	0.661	0.422	0.653	0.636	0.539	0.657
		weight	0.723	0.664	0.416	0.662	0.598	0.488	0.619
	prob.	prob.	0.739	0.675	0.430	0.672	**0.643**	0.553	0.652
		none	0.738	**0.681**	0.408	**0.678**	0.635	0.521	0.657
		weight	0.726	**0.677**	0.392	**0.680**	0.587	0.457	0.621
	weight	prob.	0.736	0.663	**0.442**	0.659	0.625	**0.571**	0.719
		none	0.740	0.667	0.435	0.663	0.632	0.550	0.720
		weight	0.730	0.672	0.427	0.672	0.593	0.498	0.720
GloVe	weight	weight	0.698	0.576	0.309	0.536	0.548	0.444	0.679

Table 4.7: Performance of different algorithms and downsampling strategies with models trained on NEWS corpus without bootstrap subsampling. **Bold** values are best or not significantly different by t-test (with $p < 0.05$). Results for SVD_{wPPMI} in *italics*.

Algorithm	Downsampling		Word Similarity				Analogy		Reliability
	Frequency	Window	MEN	MTurk	SimLex	WS-353	Google	MSR	
SVD_{PPMI}	none	prob.	0.769	0.556	0.411	0.619	0.441	0.370	0.557
		none	0.771	0.558	0.401	0.623	0.445	0.335	0.584
		weight	0.773	0.556	0.412	0.624	0.451	0.373	0.605
	prob.	prob.	0.776	0.564	0.423	0.642	0.463	0.420	0.571
		none	**0.782**	0.569	0.415	0.646	0.480	0.397	0.598
		weight	**0.782**	0.570	**0.430**	**0.654**	0.478	0.425	0.616
	weight	prob.	0.776	0.563	**0.427**	0.643	0.460	0.413	0.587
		none	**0.781**	0.571	0.417	0.645	0.476	0.390	0.617
		weight	***0.781***	*0.567*	***0.430***	***0.649***	*0.476*	*0.421*	***0.635***
SGNS	none	prob.	0.721	0.647	0.417	0.626	0.589	0.521	0.461
		none	0.726	0.661	0.408	0.630	**0.599**	0.505	0.472
		weight	0.717	0.660	0.403	0.639	0.553	0.454	0.427
	prob.	prob.	0.734	0.673	0.417	**0.647**	**0.601**	0.513	0.452
		none	0.733	**0.680**	0.394	**0.651**	0.592	0.483	0.456
		weight	0.719	**0.683**	0.377	**0.653**	0.531	0.423	0.401
	weight	prob.	0.731	0.650	**0.431**	0.638	0.577	**0.532**	0.482
		none	0.736	0.667	0.422	0.646	0.589	0.514	0.533
		weight	0.720	0.657	0.416	**0.650**	0.542	0.459	0.473
GloVe	weight	weight	0.687	0.572	0.301	0.508	0.505	0.408	0.461

Table 4.8: Performance of different algorithms and downsampling strategies with models trained on bootstrap subsampled NEWS corpus. **Bold** values are best or not significantly different by t-test (with $p < 0.05$). Results for SVD_{wPPMI} in *italics*.

4.7. Comparison of SGNS implementations

Three different implementations of SGNS were used in the previous experiments, i.e., GENSIM, HYPERWORDS and WORD2VEC.[25] These differ in how they implement randomness for vector initialization and downsampling, i.e., if they use a fixed or variable seed for random number generation (see Section 4.2). Initial word vectors are generated based on a fixed seed in HYPERWORDS and WORD2VEC and thus always identical for the same vocabulary and number of dimensions, whereas GENSIM allows for both a fixed and a variable seed. The same is true for downsampling, but HYPERWORDS can easily be modified to use a variable seed (e.g., done in Section 4.6.1). HYPERWORDS' downsampling happens in a single thread and will thus always draw the same samples from a given corpus if a fixed seed is used, whereas in GENSIM and WORD2VEC sampling is done with multiple threads, i.e., varies even with a fixed seed, as long as multiple threads are used. Also, WORD2VEC uses a separate random number generator for each thread, whereas GENSIM uses a shared one. Fixed seeds for initialization and downsampling can be expected to result in higher reliability values as they make training more deterministic. Due to the non-deterministic effect of multi-threading (see page 61), using fewer threads should also increase reliability. Using a single thread and fixed seeds should make experiments deterministic and result in perfect reliability.
Accuracy should be very similar for all implementations, but implementation details might prove to be crucial. HYPERWORDS reduces the learning rate according to the number of processed word-context pairs. In contrast, WORD2VEC reduces the learning rate based on the number of processed words. GENSIM, finally, only updates the learning rate after finishing an iteration over all examples.

4.7.1. Experimental Setup

We measure the differences in accuracy and reliability according to the general training and evaluation setup described for the NEWS corpus in Section 4.6.1, using 5 iterations and 1, 2, 5 or 10 threads for all

[25] Using Mikolov's most recent version from `https://github.com/tmikolov/word2vec` for the latter [Accessed May 28th 2019].

algorithms. The experimental code is available via GitHub.[26] Our experimental environment might influence results, as experiments were performed on a cluster[27] also providing other services.

4.7.2. Results

Tables 4.9 and 4.10 list the accuracy by similarity and analogy as well as the j@10 reliability for all three tested implementations and differing numbers of threads. We could also explore some combinations of fixed and variable seeds for GENSIM and HYPERWORDS.[28] Accuracy values for GENSIM and WORD2VEC are very similar and mirror those achieved with an identical setup in Section 4.6.2. The number of threads causes some minor changes in accuracy for almost all algorithms. However, WORD2VEC with a single thread performs worse than all other tested combinations.
Using only a single thread fixed seeds lead to perfect reliability for all three implementations (the 0.999 for HYPERWORDS is caused by one of the anchor words systematically not being processed). A high number of threads is negatively correlated[29] with reliability if at least one seed is fixed. However, this seems to be algorithm-dependent and involves some kind on non-linear saturation, i.e., there is little difference between 5 and 10 threads. The number of threads seems to have no effect if all seeds are variable, with HYPERWORDS being less reliable than GENSIM in this case. These lowest measured reliability values should be most representative for the underlying algorithm, as fixed seeds are effectively an additional parameter—recall Section 4.2 for a discussion of such fake reliability.
We deem these difference between implementations troubling, as they might mislead researchers focusing on only one. Further experiments, e.g., modifying WORD2VEC to use variable seeds, are out of scope here, since $\text{SVD}_{\text{wPPMI}}$ was already shown to be superior for our use case.

[26] `https://github.com/hellrich/sgns_implementation_comparison` [Accessed May 28$^{\text{th}}$ 2019].

[27] With 16-core® Xeon® E5-2650 v2 CPUs.

[28] Further changes to the underlying implementations would allow to test all combinations of fixed and variable seeds for each implementation, but the point of this comparison is gathering insight into these implementations as they are, not as they could be.

[29] For example, Pearson's $r = -0.66$ for WORD2VEC.

Threads	Implementation	Seed		Word Similarity				Analogy		Reliability
		Downsampling	Initialization	MEN	MTurk	SimLex	WS-353	Google	MSR	
1	GENSIM	fixed	variable	**0.740**	0.682	**0.428**	**0.669**	**0.641**	0.547	0.778
	GENSIM	fixed	fixed	**0.739**	0.681	0.427	**0.670**	**0.643**	0.546	**1.000**
	GENSIM	variable	variable	**0.739**	0.679	0.425	0.666	**0.642**	**0.550**	0.584
	HYPERWORDS	fixed	fixed	0.737	**0.693**	0.422	0.668	**0.642**	0.533	0.999
	HYPERWORDS	variable	variable	**0.741**	0.682	0.419	**0.670**	**0.643**	0.531	0.528
	WORD2VEC	fixed	fixed	0.693	0.614	0.392	0.602	0.387	0.281	**1.000**
2	GENSIM	fixed	variable	0.739	**0.683**	0.428	**0.670**	**0.641**	0.546	0.775
	GENSIM	fixed	fixed	0.739	**0.682**	0.427	**0.669**	**0.641**	0.546	**0.789**
	GENSIM	variable	variable	0.739	0.679	0.426	0.666	**0.642**	**0.550**	0.587
	HYPERWORDS	fixed	fixed	0.739	0.675	0.421	**0.670**	**0.642**	0.533	0.646
	HYPERWORDS	variable	variable	**0.743**	**0.687**	0.421	**0.667**	**0.642**	0.533	0.525
	WORD2VEC	fixed	fixed	0.738	0.673	**0.431**	0.665	**0.642**	**0.552**	0.686

Table 4.9: Performance of different SGNS implementations with fixed and variable seeds for random number generation, using either 1 thread or 2 threads. Trained on non-bootstrap subsampled NEWS. **Bold** values are best or not significantly different by t-test (with $p < 0.05$) for each number of threads.

Threads	Implementation	Seed		Word Similarity				Analogy		Reliability
		Downsampling	Initialization	MEN	MTurk	SimLex	WS-353	Google	MSR	
5	GENSIM	fixed	variable	0.739	0.675	0.425	**0.667**	**0.643**	0.549	0.584
	GENSIM	fixed	fixed	0.739	**0.678**	0.425	0.666	**0.643**	0.550	0.583
	GENSIM	variable	variable	0.739	0.677	0.425	0.668	0.642	0.549	0.578
	HYPERWORDS	fixed	fixed	0.741	**0.684**	0.418	**0.672**	0.640	0.535	0.646
	HYPERWORDS	variable	variable	**0.742**	**0.685**	0.419	**0.667**	**0.644**	0.534	0.528
	WORD2VEC	fixed	fixed	0.739	0.673	**0.428**	0.668	**0.644**	**0.554**	**0.650**
10	GENSIM	fixed	variable	0.739	**0.677**	0.425	0.652	**0.644**	**0.549**	0.588
	GENSIM	fixed	fixed	0.739	**0.676**	0.425	0.651	**0.643**	**0.549**	0.589
	GENSIM	variable	variable	0.738	0.672	0.425	**0.652**	**0.644**	0.547	0.587
	HYPERWORDS	fixed	fixed	**0.742**	**0.682**	0.419	**0.654**	0.640	0.536	0.634
	HYPERWORDS	variable	variable	**0.742**	**0.681**	0.422	**0.656**	**0.643**	0.533	0.524
	WORD2VEC	fixed	fixed	0.739	**0.680**	**0.428**	0.651	**0.643**	**0.550**	**0.637**

Table 4.10: Performance of different SGNS implementations with fixed and variable seeds for random number generation, using either 5 or 10 threads. Trained on non-bootstrap subsampled NEWS. **Bold** values are best or not significantly different by t-test (with $p < 0.05$) for each number of threads.

4.8. Discussion

Overall, our experiments lead to a troublesome conclusion: Neither SG(NS) nor GloVe can provide reliable word embeddings. They are thus ill-suited for an increasingly popular form or corpus linguistic research, where word meaning is determined via most similar words. This can be seen as part of a general reproducibility crisis affecting artificial intelligence research, which hampers comparisons and thus stifles progress (Henderson et al., 2018; Hutson, 2018; Reimers & Gurevych, 2017).

Luckily, singular value decomposition based word embeddings—and especially our novel SVD_{wPPMI}—are a reliable solution for corpus linguistic studies. Their results are highly reproducible, resulting in identical embeddings for repeated training on one corpus. They are still at least as reliable as alternative embedding methods if corpora are modified by bootstrap subsampling. Reliability on large bootstrap subsampled corpora approaches that of other embedding methods on un-modified corpora. Reliability problems could also be mitigated by using some kind of ensemble (Antoniak & Mimno, 2018), as we already briefly suggested in Hellrich & Hahn (2016b), however this would lead to increased computational demands.

Word embedding reliability has, to the best of our knowledge, only recently become of interest to other researchers.[30] Few studies are comparable with our own work:

Antoniak & Mimno (2018) used an approach similar to Hellrich & Hahn (2017a) and compared SGNS, GloVe, LSA (Deerwester et al., 1990) and a form of SVD_{PPMI} (using probabilistic downsampling with a variable seed) models with j@20. Their assessment was based on three domain specific corpora (news paper articles, legal texts and internet forums) and a large number of models (50), but only a small number of anchor words (20). They used not only their corpora as-is, but also shuffled and bootstrap subsampled versions. They found GloVe to be most reliable and SVD_{PPMI} to be worst—as we showed in in Section 4.6 their assessment of SVD_{PPMI} is only due to their downsampling strategy.

[30] Several studies use 'stability' instead of 'reliability'. Both differ in their focus on the variance of vector representations or the resulting variance in similarity judgments, respectively.

Wendlandt et al. (2018) compared embeddings trained with different algorithms, parameters (e.g., 50–800 dimensions) and even corpora with a r@10 metric. Their work is thus important for the general study of word embeddings, but less relevant for their suitability as a corpus linguistic tool. They compared SGNS, GloVe and a proto form of SVD_{PPMI} embeddings (Bullinaria & Levy, 2007), using corpora from two domains, i.e., newspaper articles and parliament records. With the 2.5k words present in all corpora as anchor words, they found GloVe to be most reliable, closely followed by their SVD based embeddings. Factors with a high impact on reliability were the order in which training examples were processed, the part-of-speech of an anchor word and the domain a model was trained on. In contrast to our results from Sections 4.3 and 4.4 they found word frequency to have little impact on reliability—this might be due to a probably overall high frequency of their anchor words, which had to be present in all corpora.

Pierrejean & Tanguy (2018a) used three corpora, one balanced and two from the scientific domain, to assess the reliability of SGNS with a variant of r@n. They used 5 models and all adjectives, nouns and verbs with at least 100 occurrences as anchor words. A novel result is their identification of clusters of words with high reliability.

Finally, Chugh et al. (2018) used three relatively small corpora to investigate the effect of embedding dimensionality on reliability.

Detailed comparisons between these studies and our own results are hampered by differences in hyperparameters (such as dimensionality), anchor word selection and corpus choice. Most notably, all corpora used in other reliability studies were relatively small, ranging from only 1.2M tokens in Chugh et al. (2018)[31] to 100M in Pierrejean & Tanguy (2018a). However, despite these differences, all studies agree on the existence of a word embedding reliability problem.

There are some additional studies touching the reliability of word embeddings which are less directly comparable, but still relevant. Heimerl & Gleicher (2018) explored novel visualizations for comparing different representations of a word due to language change or lacking reliability. Fares et al. (2017) suggested creating repositories of word embeddings for use in subsequent studies to increase comparability. While we welcome this idea in general—and provide such a repository ourselves for the vectors used in JeSemE (see Section 5.2)—it would

[31] Size information from personal communication.

just mask the arbitrary nature of SGNS and GloVe embeddings and lead to the problems discussed for fixed seed values in Section 4.2. Tahmasebi (2018) used j@10 in a study on word change in a Swedish historical newspaper collection. They trained SGHS models and compared the most similar words for 11 manually selected anchor words between successive years. They thus trained only one model per sub-corpus and not multiple as in the studies above. They caution using word2vec, especially for studies interested in finer details such as sense change. Pierrejean & Tanguy (2018b) explored the impact of different corpora and parameters (such as dimensionality) on the most similar words provided by SGNS and CBOW word embeddings. They used only one model per corpus-parameter combination and are thus likely to measure artifacts caused by inherently low reliability—which is strange given their work on this phenomenon in Pierrejean & Tanguy (2018a). Finally, Dubossarsky et al. (2017) investigated the validity of several high-ranking publications investigating language change in general as already discussed in Section 2.5.3.

Overall, we believe it prudent to use $\text{SVD}_{\text{wPPMI}}$ in embedding-based corpus linguistic research wherever possible, i.e., except for data sets where the resulting word context matrix is too large and a streaming approach must be used. In those cases, and also if some kind of random sampling is performed, we recommend providing some form of reliability measurement. Given the results of Dubossarsky et al. (2017) it might also be interesting to explore non-word embedding approaches, but most studies show these to under-perform on semantic tasks. Further methodological studies as well as case-studies on specific words seem currently more prudent than large scale investigations on lexical change in general (see Section 2.5.3).

Chapter 5

Observing Lexical Semantic Change

This chapter contains our[1] work on lexical semantic change affecting both denotation and emotional connotation.
Section 5.1 describes joint work with Sven Buechel[2] on modeling historical word emotions. Its results were partially presented at the LT4DH workshop (Buechel et al., 2016), the Digital Humanities conference (Buechel et al., 2017) and the LaTeCH-CLfL workshop (Hellrich et al., 2019a).
Section 5.2 describes JeSemE, the Jena Semantic Explorer, a website for visualizing semantic change based on several diachronic corpora. It was previously presented at ACL and COLING (Hellrich et al., 2018a; Hellrich & Hahn, 2017b). It also utilizes results from Section 5.1 to model word emotions.
Finally, Section 5.3 contains two case studies using JeSemE. One of these case studies, i.e., joint work with Alexander Stöger[3] on investigating the history of science with distributional methods, was previously presented at the DHd conference (Hellrich et al., 2018b).

[1] Plural is used as experiments in this chapter were planned and published in co-operation with my supervisor Professor Dr. Udo Hahn and in some cases also Sven Buechel or Alexander Stöger (see next footnotes for details).

[2] The general idea of predicting historical word emotions as well as the adaptation of algorithms and the creation of test sets were developed jointly. Background knowledge on word emotion modeling was provided by Sven Buechel, who also conducted the analyses of emotions in texts from different domains and points in time. Corpora and parameters for word embeddings were selected by me.

[3] Alexander Stöger provided background knowledge on the investigated historical development as well as expert judgment on the validity of our results.

We provide thus both a novel method and an easy-to-use tool for diachronic studies. Our case studies demonstrate possible applications and their results indicate JeSemE to produce valid, or at least plausible, results.

5.1. Historical Word Emotions

As discussed in Section 2.5, the meaning of words consists of both denotation and connotation. Word embeddings do model both (Rothe et al., 2016), but do not provide explicit access to either of them. Research in psychology and computational linguistics makes it possible to make the emotional connotation explicit.
This is highly relevant for studies in the (digital) humanities which try to track emotions in historical texts. For example Acerbi et al. (2013) and Bentley et al. (2014) observed long term trends in words expressing emotions in the Google Books corpus and could link those to historical (economical) events, but their study relied on contemporary word emotion information (and might also be influenced by sampling problems in the underlying corpus (Pechenick et al., 2015)). Another example is Kim et al. (2017) investigating emotions in literary texts to find genre-specific patterns while also relying on (too) contemporary word emotion information.
Our methods for temporal adaption are based on methods for expanding word emotion lexicons. We were the first to perform such an adaption with a fine-grained dimensional model (see below). Our adaption was also the first to be evaluated against human expert judgment, providing an indication of its validity. It was also the first to operate on German data.

5.1.1. Related Work

Quantitative models for word emotions can be traced back at least to Osgood (1953), who used questionnaires to gather human ratings for words on a wide variety of dimensional axes including good–bad. We use a dimensional model with three axes, i.e., Valence-Arousal-Dominance (VAD; Bradley & Lang (1994)) as illustrated in Figure 5.1. In this model Valence encodes whether an emotion is positive or negative, e.g., joy being more positive than fear, Arousal encodes whether an emotion is connected with calmness or excitement, e.g.,

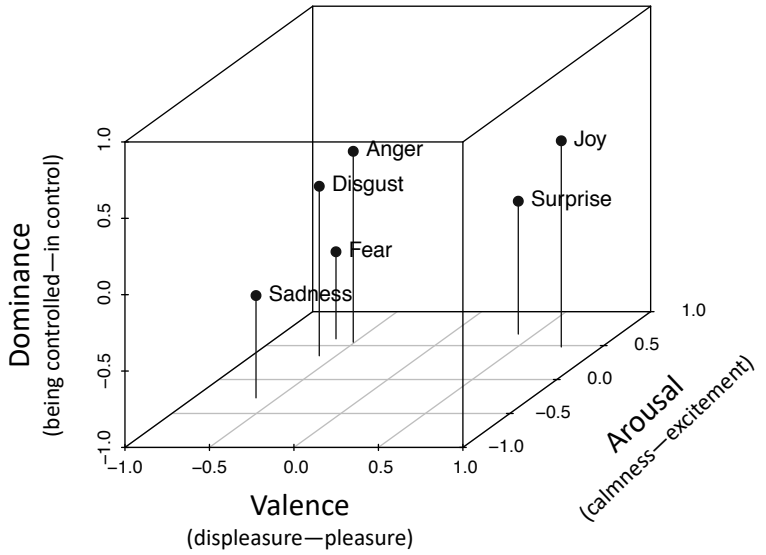

Figure 5.1: VAD emotion space with position of basic emotions (Ekman, 1992) for reference. Adapted from Buechel & Hahn (2016).

sadness being more calm than anger, and Dominance encodes the perceived degree of control, e.g., fear making one feel less in control than joy. Using categories, such as anger or sadness in the previous examples, is an alternative approach already employed in the General Inquirer text analysis system (Stone & Hunt, 1963). The so-called basic emotions (Ekman, 1992) form a prominent categorical model. They are defined as cultural universals with links to evolutionary biology (e.g., unique facial expression, present in other primates). However, there is no agreement on which emotions are basic emotions under these criteria, with Ekman (1992, p. 193) listing twelve strong candidates and several more that might be covered by a slightly relaxed definition. Polarity, an even simpler emotion format, is very popular for applications known as sentiment analysis or opinion mining (Pang & Lee, 2008; Turney, 2002). Here only two categories—Positive and Negative—or a single positive–negative axis (corresponding to VAD's Valence dimension (Calvo & Kim, 2013)) is used. The three-dimensional VAD is thus both extensive and elegant,

Word	Valence	Arousal	Dominance
rage	2.50	6.62	4.17
orgasm	8.01	7.19	5.84
relaxed	7.25	2.49	7.09

Table 5.1: Sample Valence-Arousal-Dominance (VAD) ratings from the emotion lexicon by Warriner et al. (2013). The scales span the interval of $[1, 9]$ for each dimension, "5" being neutral.

especially as high-quality mappings between different representations are possible when necessary (Buechel & Hahn, 2018a).

Word emotion information is collected (mainly by psychologists) in word emotion lexicons—see Table 5.1 for an illustration of the structure of such a lexicon. For English, the *Affective Norms of English Words* (ANEW; Bradley & Lang (1999)) incorporate 1,034 words paired with experimentally determined affective ratings using a 9-point scale for Valence, Arousal and Dominance, respectively. Warriner et al. (2013) provided an extended version of this resource (14k entries) employing crowdsourcing. As far as German-language emotion lexicons are concerned, ANGST (Schmidtke et al., 2014) is arguably the most important one for NLP purposes. It comprises 1,003 lexical entries and replicates ANEW's methodology very closely (see Köper & Schulte im Walde (2016) for a more complete overview of German VAD resources).

We aim to predict VAD values with a regression task to allow for a fine-grained analysis. We do this by adapting algorithms for the expansion of emotion lexicons. Such algorithms use information on word similarity (or even position in vector space) and a limited amount of seed words with known emotions to predict emotions connected with arbitrary words. More than a decade ago, Turney & Littman (2003) introduced an algorithm which was frequently used or adapted by others (Köper & Schulte im Walde, 2016; Palogiannidi et al., 2016). It computes a sentiment score based on the association (by PMI) of an unrated word to two sets of positive and negative seed words, respectively. Bestgen (2008) presented an algorithm which has been prominently put into practice to expand an existing VAD lexicon (Bestgen & Vincze, 2012). Their method employs a k-Nearest-

Neighbor methodology where an unrated word inherits the averaged rating of the surrounding words. Rothe et al. (2016) presented a more recent approach to polarity induction. Based on word embeddings and a set of positive and negative paradigm words, they train an orthogonal transformation of the embedding space so that the encoded polarity information is concentrated in a single vector component whose value then serves as an explicit polarity rating. The algorithm proposed by Hamilton et al. (2016a) employs a random walk within a lexical graph constructed using word similarities. They outperformed Rothe et al. (2016) in a comparison on small corpora and consider their method especially suited for historical applications.[4]
Algorithms for bootstrapping word emotion information can also be used to predict historical emotion values by using word similarity based on historical texts. This was first done for polarity regression with the Turney & Littman (2003) algorithm and a collection of three British English corpora by Cook & Stevenson (2010). Jatowt & Duh (2014) tracked the emotional development of words by averaging the polarity of the words they co-occurred with (assuming the latter's polarity to be stable). Hamilton et al. (2016a) used their novel random walk based algorithm for polarity regression on COHA. This algorithm was also used by Généreux et al. (2017) to test the temporal validity of inferred word abstractness, a psychological measure akin to the individual VAD dimensions. They used both modern and historical (1960s) psychological datasets rating the same words as gold standards and found a strong correlation with predicted historical abstractness.

5.1.2. Historical Emotion Gold Standard

In general, native speakers are the best option for acquiring a gold standard lexicon of emotional meaning for any language or domain. These are obviously not available for most historic varieties due to the limited human lifespan (see also Section 2.5.4. We instead rely on historical language experts for constructing our data set. The gold standard consists of two parts, an English and a German one, with 100 words each. We recruited three annotators for German and two

[4] However, the algorithm is sensitive to changes in its training material and thus likely prone to artifacts, see README of `https://github.com/williamleif/socialsent` [Accessed May 28th 2019].

for English, all doctoral students (four at the graduate school "The Romantic Model") and experienced with interpreting 19th century texts.

We selected high frequency words for the annotation to ensure high quality of the associated word embeddings. The selection was done by extracting adjectives, common nouns and lexical verbs from the 1830s COHA and the 1810–1839 DTA subcorpus. We then randomly sampled 100 words out of the 1000 most frequent ones.

The rating process was set up as a questionnaire study following designs from psychological research (Bradley & Lang, 1999; Warriner et al., 2013). The participants were requested to put themselves in the position of a person living between 1810 and 1839 (German) respectively in the 1830s (English). They were then presented with stimulus words and used the so-called self-assessment manikin (SAM) to judge the kind of feeling evoked by these lexical items (Bradley & Lang, 1994). SAM consists of three individual nine-point scales, one for each VAD dimensions. Each of the 27 rating points is illustrated by a cartoon-like anthropomorphic figure in addition to verbal anchors for the low and high end of the scales, e.g., the rating point "9" of the Valence scale representing "complete happiness". The final ratings for each word were derived by averaging the individual ratings of the annotators.

We measured inter-annotator agreement (IAA) by calculating the standard deviation for each word and dimension, thus constituting an error-based score (lower is better). Deviations were averaged for each individual VAD dimension first and then averaged again over these aggregate values.

Table 5.2 shows the resulting IAA for each dimension. In comparison with the lexicon by Warriner et al. (2013), our gold standard displays higher rating consistency. This suggests that experts show higher agreement, even when judging word emotions for a historical language stage, than crowdworkers for contemporary language. The resulting gold standards are available online.[5]

[5] `https://github.com/JULIELab/HistEmo/tree/master/historical_gold_lexicons` [Accessed May 28th 2019].

	Valence	Arousal	Dominance	Average
goldEN	1.20	1.08	1.41	1.23
goldDE	1.72	1.56	2.31	1.86
Warriner	1.68	2.30	2.16	2.05

Table 5.2: Inter-annotator agreement for our English (goldEN) and German (goldDE) gold standard, as well as the lexicon by Warriner et al. (2013)—for comparison—for each VAD dimension, as well as averaged over the three dimensions.

5.1.3. Algorithms and Adaptations

Our work employs three algorithms for inducing emotion lexicons, two of which had to be adapted to deal with the more informative vectorial VAD representation instead of a binary (positive vs. negative polarity) representation:

kNN The k-Nearest-Neighbor-based algorithm by Bestgen (2008) which already supports vectorial input.

ParaSimNum An adaptation of the ParaSim algorithm by Turney & Littman (2003) which is based on the similarity of two opposing sets of paradigm words and ill-suited for vectorial input.

RandomWalkNum An adaptation of the RandomWalk algorithm proposed by Hamilton et al. (2016a) which propagates affective information of seed words via a random-walk through a lexical graph. It cannot process vectorial input.

kNN sets the emotion values $\hat{e}(w)$ of each word w to the average of the emotion values of the n most similar seed words. Let $e(s)$ map a seed word to a three-dimensional vector corresponding to its VAD value in our seed lexicon and $\mathit{nearest}(w, n)$ map to a set of the n most similar seed words s for a given word w:

$$\hat{e}_{kNN}(w, k) := \frac{1}{k} \sum_{s \in \mathit{nearest}(w,k)} e(s) \tag{5.1}$$

ParaSim computes $\hat{e}(w)$ by using the function $sim(w, p)$ to measure the similarity of a word w with a set of positive and negative paradigm words (POS and NEG, respectively):

$$\hat{e}_{ParaSim}(w) := \sum_{p \in POS} sim(w, p) - \sum_{n \in NEG} sim(w, n) \tag{5.2}$$

Let $e(s)$ map to '1' if word $s \in POS$ and to '-1' if $s \in NEG$, then Equation 5.2 can be rewritten as:

$$\hat{e}_{ParaSim}(w) := \sum_{s \in POS \cup NEG} sim(w, s) \times e(s) \tag{5.3}$$

The adaptation to numerical input in ParaSimNum is achieved by changing e(s) to map to a three-dimensional VAD vector for each word in our seed lexicon L and normalizing the resulting value:

$$\hat{e}_{ParaSimNum}(w) := \frac{\sum_{s \in L} sim(w, s) \times e(s)}{\sum_{s \in L} sim(w, s)} \tag{5.4}$$

RandomWalk propagates sentiment scores through a graph, with vertices representing words and edge weights denoting word similarity. The vector $p \in \mathbb{R}^{|\mathcal{V}|}$ ($\mathcal{V}$ being the set of words which make up the lexical graph) represents the induced sentiment score for every word in the graph. It is iteratively updated by applying the transition matrix T (see Zhou et al. (2003) for more details):

$$p^{(t+1)} = \beta T p^{(t)} + (1 - \beta) s \tag{5.5}$$

Here $s \in \mathbb{R}^{|\mathcal{V}|}$ is the vector representing the seed sentiment scores and the β-parameter balances between assigning similar scores on neighbors and correct scores on seeds. The vector p is initialized so that the i-th element $p_i = 1/|\mathcal{V}|$, whereas s is initialized with $s_i = 1/|\mathcal{S}|$ ($\mathcal{S}$ being the set of seed words), if the corresponding word w_i is a seed word and 0, otherwise.
To obtain the final sentiment scores p_{final}, the process is independently run until convergence for both a positive and a negative seed set, before the resulting values p^+ and p^- are normalized by performing a z-transformation on:

$$p_{final} := \frac{p^+}{p^+ + p^-} \tag{5.6}$$

For RandomWalkNum, p and s are replaced by $|\mathcal{V}| \times 3$ matrices, P and S. All entries of P are initialized with $1/|\mathcal{V}|$. For the positive seed set, S is populated with the original VAD values of each word in the seed lexicon and 0, otherwise. For the negative seed set all values are inverted relative to the center of the numerical VAD rating scales, e.g., for the examples in Table 5.1 the valence score of *relaxed* is transformed from 7 to 3. In both cases S is then normalized so that each column adds up to 1. P_{final} can than be calculated like p_{final} in the original algorithm.

5.1.4. Experimental Setup

Our experiments were intended to compare combinations of three algorithms, i.e., kNN, ParaSimNum and RandomWalkNum, and embedding types, i.e., SVD_{PPMI} and SGNS. We also explored the effect of seed lexicon size, as Hamilton et al. (2016a) used only a very short list, whereas Cook & Stevenson (2010) used a large one.
The word embedding training follows Hamilton et al. (2016a). COHA and DTA were preprocessed by using the lemmatization provided with each corpus, removing punctuation and converting to lower case. We then used hyperwords (Levy et al., 2015) to create both a SVD_{PPMI} and an SGNS model[6] for three temporal slices, i.e., 1810–1839 DTA as well as 1830s and 2000s COHA.
We tested kNN, ParaSimNum and RandomWalkNum both in a synchronic and in a diachronic scenario. The synchronic scenario, i.e., reconstructing Warriner's contemporary emotion lexicon with embeddings trained on recent language (2000s COHA), provides an upper bound on algorithmic performance. In contrast, the diachronic scenario tests directly how well predictions with embeddings trained on historical language match our historical emotion gold standard. For English, we used two different seed lexicons. The full seed lexicon corresponds to all the entries of words which are present in Warriner's VAD lexicon and Anew (about 1,000 words). In contrast, the limited

[6] We used 300 dimensions, a context window of up to four words (limited by document boundaries, but ignoring sentence boundaries) and a minimum word frequency threshold of 100. Eigenvectors were discarded and no negative sampling was used for SVD_{PPMI}. Word and context vectors were combined to create the final embeddings.

seed lexicon is restricted to the 19 words[7] which were identified as temporally stable by Hamilton et al. (2016a). For German, we tested only a full seed lexicon, ANGST, as most entries of the English limited lexicon have no corresponding entries in ANGST. Our evaluation uses Pearson's r between actual and predicted values for each affective dimension (Valence, Arousal and Dominance) for quantifying performance. Our values are thus not comparable with those of Généreux et al. (2017), who used the rank based Spearman's ρ.[8]

5.1.5. Results

Table 5.3 provides correlation (Pearson's r) averaged over all VAD dimension[9] each seed lexicon, embedding method and induction algorithm for our synchronic experiment. SGNS embeddings are worse than SVD embeddings for both full and limited seed lexicons. SVD_{PPMI} embeddings seem to be better suited for induction based on the full seed set, leading to the highest observed correlation with PARASIMNUM. However, differences to the other algorithms are statistically non-significant. Conversely, the results are clearer using the limited seed set. Here, RANDOMWALKNUM is significantly better than all alternative approaches, but results are also far worse than those with the full seed lexicon.
Table 5.4 provides the average values of these VAD correlations for each seed lexicon, embedding method and induction algorithm for our diachronic experiment. For English using the full seed lexicons, we found $r \approx .35$, with no significant difference between the different approaches due to the small size of the gold standard. Notably, the limited seed lexicon performed markedly weaker in every single condition. This finding directly contradicts the claim by Hamilton et al. (2016a) that small temporally stable seeds words are preferable over larger and thus noisier ones. Results for German (using the full

[7] One of the 20 words given by Hamilton et al. (i.e., *hated*) is not present in the Warriner lexicon and was therefore omitted.

[8] Some other studies on emotion lexicon expansion also used Kendall's τ. We found τ values to be consistent with r values during a pretest.

[9] Performance is known to differ between VAD dimensions, i.e., Valence is usually the easiest one to predict. For the full seed lexicon and the best induction method, PARASIMNUM with SVD_{PPMI} embeddings, we found correlation values between 0.679 for Valence, 0.445 for Arousal and 0.547 for Dominance.

seed lexicon) are similar to those for English. However, the SNGS embeddings are here outperformed by the SVD_{PPMI} ones.
Our most important empirical result is that limiting seed words to supposedly temporally stable ones does not improve performance as suggested by Hamilton et al. (2016a) but instead turns out to be harmful. Overall, we deem using ParaSimNum with SVD_{PPMI} and full seed lexicons to be the best solution, as its results are at least competitive and no further parameters (e.g., a number of next neighbors) must be chosen.

Induction Method	Seed Lexicon	SVD_{PPMI}	SGNS
kNN	full	**0.548**	0.487
ParaSimNum	full	**0.557**	0.489
RandomWalkNum	full	**0.544**	0.436
kNN	limited	0.181	0.166
ParaSimNum	limited	0.249	0.191
RandomWalkNum	limited	**0.330**	0.181

Table 5.3: Results of the synchronic evaluation in Pearson's r averaged over all three VAD dimensions. Best system for each seed selection strategy (full vs. limited) and those with non-significant differences ($p \geq 0.05$) in **bold**.

Language	Induction Method	Seed Lexicon	SVD_{PPMI}	SGNS
English	kNN	full	0.307	**0.365**
	ParaSimNum	full	0.348	0.361
	RandomWalkNum	full	0.351	0.361
	kNN	limited	0.273	0.153
	ParaSimNum	limited	0.295	0.232
	RandomWalkNum	limited	**0.305**	$0.039^{\triangle}$
German	kNN	full	0.366	0.263
	ParaSimNum	full	**0.384**	0.214
	RandomWalkNum	full	0.302	0.273

Table 5.4: Results of the diachronic evaluation in Pearson's r averaged over all three VAD dimensions. The best system for each language and seed selection strategy (full vs. limited) is in **bold**, however only the system marked with '$^{\triangle}$' is significantly different ($p < 0.05$).

5.2. JeSemE — Jena Semantic Explorer

JeSemE[10] is an open source[11] website for investigating semantic change with distributional methods. It makes state-of-the-art methods accessible to all linguists and scholars in the (digital) humanities, as it requires neither technical skills nor computational resources from its users. JeSemE can be queried for semantic change in five diachronic corpora, i.e., COHA, DTA, GBF, GBG and RSC (see Chapter 3). JeSemE provides a plethora of information for each word queried:

- Most similar words over time modeled with SVD_{PPMI}.[12]
- Changes in word emotions modeled with the ParaSimNum algorithm (see Section 5.1).
- Strongly associated words as tracked with two word association measures, i.e., PPMI and χ^2.
- Information on relative word frequency over time.

JeSemE is superior to alternative systems in both capabilities and corpus coverage. This section describes JeSemE's resources, architecture and interface, as well as alternative systems. Several detailed use cases based on JeSemE's current version 2.1—the underlying models are also stored on Zenodo[13] to ensure long-term availability—can be found in Section 5.3.

5.2.1. Used Corpora

To achieve sufficient training data with a consistent lower bound on size we divided our corpora in temporal slices with a minimum size of about 10M tokens or 5-grams. These cover 10 years each for COHA, GBF and GBG, as well as 30 years for the smaller DTA and finally two 50 year slices and one 19 year[14] slice for the even smaller

[10] `http://jeseme.org` [Accessed May 28th 2019].

[11] `https://github.com/JULIELab/JeSemE` [Accessed May 28th 2019].

[12] $\text{SVD}_{\text{wPPMI}}$ (see Section 4.6) was developed after most studies described in this chapter were conducted and is not yet integrated in JeSemE.

[13] `https://doi.org/10.5281/zenodo.2605352` [Accessed May 28th 2019].

[14] This slice is about equal in size to its two predecessors despite the lower amount of years.

RSC. Table 5.5 shows the resulting effective corpus size as well as the number of types which could be modeled consistently, i.e., belong to the 10k (5k for RSC) most frequent in each temporal slice.[15] Note that JeSemE does not cover the whole time span covered by each corpus as described in Chapter 3, since some temporal slices contained not enough tokens or 5-grams and were thus excluded.

Corpus	Years	Slices	Types	Tokens	5-grams
COHA	1830–2009	18	5,101	376M	–
DTA	1751–1900	5	5,347	81.0M	–
GB Fiction	1820–2009	19	6,492	–	14.7G
GB German	1830–2009	18	4,450	–	5.25G
RSC	1750–1869	3	3,080	24.7M	–

Table 5.5: Details on corpora as used in JeSemE, i.e., covered years, number of temporal slices, number of modelled types and overall size in number of tokens or 5-grams (G is shorthand for 10^9).

5.2.2. System Architecture

JeSemE has two main components, a pipeline for processing corpora to derive semantic information and a web service for providing this information to users. Its pipeline utilizes a modified[16] version of hyperwords (Levy et al., 2015), while the web service is built with the Spark Web framework[17] and runs inside a Jetty[18] web server. Semantic information is stored in a PostgreSQL[19] database which thus links the two main components.

[15] In honor of my graduate school an exception was made for the word *Romantik* ['romantic', noun] in GBG which was slightly less frequent for one time span.

[16] `https://github.com/hellrich/hyperwords` [Accessed May 28th 2019].

[17] Not to be confused with the Apache Spark Big Data framework; see: `http://www.sparkjava.com` [Accessed May 28th 2019].

[18] `http://www.eclipse.org/jetty` [Accessed May 28th 2019].

[19] `https://www.postgresql.org` [Accessed May 28th 2019].

JeSemE's four step processing pipeline is illustrated in Figure 5.2:

- Firstly, non-alphanumeric characters are removed from the five raw corpora and the English corpora are converted to lower case. Normalized text is already provided in DTA, whereas GBG is processed as described in Section 4.4.1.

- Secondly, our modified HYPERWORDS is used to calculate χ^2 and PPMI association scores as well as SVD_{PPMI} embeddings for each temporal slice.[20] PPMI and were χ^2 were normalized, so that all association scores for a given word add up to one—this step is intended to make the resulting values easier to compare.

- Thirdly, we extrapolate historical word emotion values by combining present day word emotion lexicons (Schmidtke et al., 2014; Warriner et al., 2013) with the word embeddings from the previous step. The lexicon expansion algorithm from Turney & Littman (2003) is used for this step. Details on word emotion modeling are provided in Section 5.1.

- Finally, we store the information derived in the previous steps, i.e., word embeddings, word association, word frequency and word emotions, in our database. The JeSemE version presented at the ACL 2017 conference stored pre-computed similarity values between all words, whereas our latest version uses word embeddings to calculate similarity on the fly—this reduces database memory demands from approximately 120GB to 40GB. The most similar words for each word (used as reference in visualizations, see below) are cached for faster retrieval.

5.2.3. User Interface

The interactive JeSemE website consists of four pages, i.e., a `search` page, a `result` page, a `help` page and an `about` page (providing legally required information). The website is built, among several

[20] We used 500 dimensions, a 4 word context window (maximum size possible with Google corpora), no downsampling based on high frequency or distance, context distribution smoothing with $\alpha = 0.75$ and a minimum frequency threshold of 100 for Google Books corpora and 50 for others.

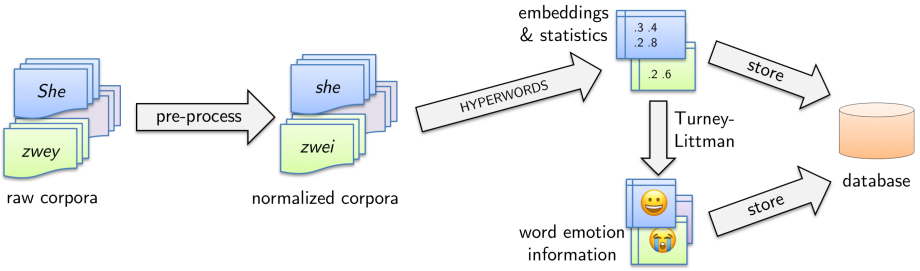

Figure 5.2: Diagram of JeSemE's processing pipeline.

other standard components, with the Thymeleaf[21] template engine and the C3[22] visualization framework.

The `search` page, shown in Figure 5.3, allows users to choose a corpus and word under scrutiny, the latter being automatically lowercased or lemmatized as necessary.

The `result` page then provides four interactive line plots[23] showing temporal trends in the most similar words, emotion values (in standard deviations from the average of all words), typical context according to association (can be toggled between χ^2 and PPMI, both normalized) as well as word frequency. Most similar words and typical contexts start with up to four reference words, with two each selected for being most similar/associated in the first, respectively last, time period modeled for this corpus. Additional reference words can be chosen with a small search bar next to each plot, thus allowing for arbitrary comparisons. Figure 5.4 shows a cropped screenshot of the `result` page for querying *heart* in COHA.

Our choice of line plots for the most similar words follows Kim et al. (2014). While it might be slightly counterintuitive not to show the word under scrutiny as part of the plot, it allows us to use a single plot style for all offered information. We refrain from using a two-dimensional projection for visualization (used in e.g., Kulkarni et al. (2015), Hamilton et al. (2016c)), as we deem it potentially

[21] `https://www.thymeleaf.org` [Accessed May 28th 2019].

[22] `http://c3js.org` [Accessed May 28th 2019].

[23] Bar plots are used for all visualizations involving RSC as it has only three temporal slices.

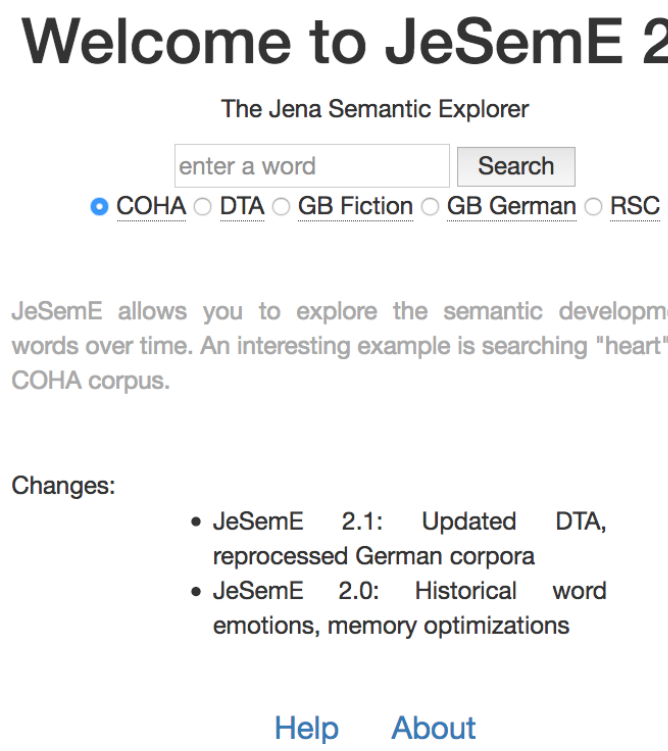

Figure 5.3: Screenshot of JeSemE's `search` page.

misleading by implying a constant meaning of those words used as the background (which are actually positioned by their meaning at a single point in time). A more suitable alternative could be the novel visualization style developed by Heimerl & Gleicher (2018) further discussed in Section 5.2.4.

In addition to the interactive website, we also provide a REST API. API calls need to specify the corpus to be searched and one (frequency, emotions) or two (similarity, context) words as `GET` parameters,[24] details being described on JeSemE's `help` page.[25] This page also contains short descriptions of the methods and corpora used in JeSemE as well as references to relevant publications.

[24] For example: `http://jeseme.org/api/similarity?word1=Tag&word2=Nacht&corpus=dta` [Accessed May 28th 2019].

[25] `http://jeseme.org/help.html#api` [Accessed May 28th 2019].

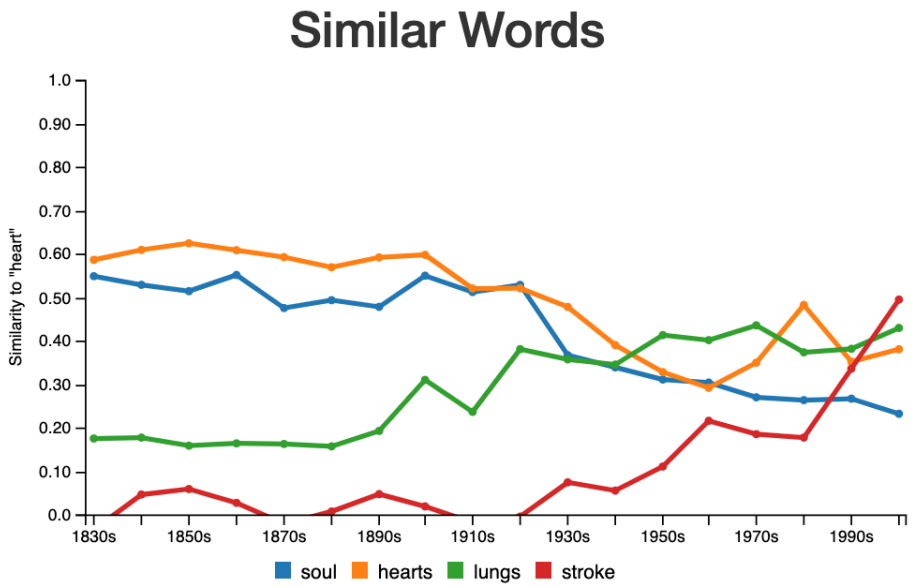

Figure 5.4: Screenshot of JeSemE's `result` page when searching for *heart* in COHA, cropped to highlight the most similar word results.

5.2.4. Alternatives

Several alternative websites[26] for exploring semantic change exist. However, none of these offers a similar combination of corpus coverage and functionality. Due to our focus on semantic change, websites aimed at syntactic change (e.g., Schätzle et al. (2017)'s HistoBankVis for Icelandic) are out of scope. To avoid copyright issues, no screenshots are provided.

[26] Locally installed tools would be ill-suited for word embedding based diachronic studies by non-technical users due to computational demands and long training times.

Davies provides a website[27] for exploring a wide range of corpora, including COHA and some Google Books corpora, but not GBF and GBG (Davies, 2012, 2014). Besides providing information on diachronic frequency, it also shows specific context words (sorted by frequency or a form of pointwise mutual information) over time. Lists of (context) words or tables containing context words for a specific time span are used for visualization. No API is provided[28] and users need to create an account to perform more than a small amount of queries. Comparisons between words are supported for COHA, however consist only of two lists of collocates displayed next to each other.

DTA can be queried online with DiaCollo[29] for word frequency and specific context words according to several association metrics (Jurish, 2015). DiaCollo offers multiple visualizations, e.g., line and bubble charts, as well as JSON or CSV for download. Words can be compared both by their specific context words and frequency.

ESTEEM[30] is intended to track changes in word similarity in social media based on word embeddings (Arendt & Volkova, 2017). Corpus selection is very limited and ill-suited for studies in diachronic linguistic (two sets of tweets collected in 2016). They use a simple yet elegant visualization, showing all words that were among the most similar words for a query word on the y-axis while the x-axis indicates whether a word was among the most similar words at a specific point in time.

Heimerl & Gleicher (2018) created a novel visualization for word embedding derived similarity over time. In their online demo[31] words on the left side of the x-axis are more similar to a query word than those on the right and words can be selected to highlight their development over time. The search interface of the demo is rudimentary and they re-used SGNS embeddings from another study (Hamilton et al., 2016c).

[27] `https://corpus.byu.edu` [Accessed May 28th 2019].

[28] `https://corpus.byu.edu/faq.asp#x10a` [Accessed May 28th 2019].

[29] `http://kaskade.dwds.de/dstar/dta/diacollo` [Accessed May 28th 2019].

[30] `https://esteem.labworks.org` [Accessed May 28th 2019].

[31] `http://embvis.flovis.net/s/neighborhoods.html` [Accessed May 28th 2019].

Gamallo et al. (2018) created the DIACHRONIC EXPLORER,[32] a website for tracking diachronic changes in the Spanish Google Books corpus. They calculate word similarity by representing each word with a sparsified (syntactic) context vector, i.e., containing zeroes as association for all contexts, except for a small number of highly associated ones. Like JESEME, they use line plots for visualization. Finally, Li et al. (2019) created the MACROSCOPE,[33] which is overall very similar to JESEME in its capabilities. It provides trends in both word similarity (based on SVD_{PPMI}) and emotion with a variety of visualizations, e.g., line plots and clustering. Word emotions are represented with a four dimensional model, i.e., concreteness in addition to VAD, and inferred with a simple algorithm averaging the emotion values of co-occurring words. Corpus selection is currently limited to the English Google Books corpus.

5.3. Insights for the Digital Humanities

The following case studies demonstrate the insights available with JESEME and its ability to track changes in general word meaning and word emotion. We used examples which are relevant for the (digital) humanities, i.e., history of science and literary studies. The aim of this section is to demonstrate possible applications and not an in-depth study of the underlying questions, which would be widely out of scope. Due to the qualitative and ex post nature of the following interpretations, they can only indicate a general aptness of JESEME—given the general lack of evaluation options discussed in Section 2.5.4, they should nevertheless be insightful. The experimental setup of all studies consists mostly of querying JESEME (see Section 5.2) for words of interest. Additional reference words were added (through JESEME's interface) as necessary. In some cases we also looked directly at the underlying corpora to provide text examples[34] for further clarification. Note that JESEME operates on lower case words for English, respectively normalized words for German.

[32] `https://github.com/citiususc/explorador-diacronico` [Accessed May 28th 2019].

[33] `http://www.macroscope.tech` [Accessed May 28th 2019].

[34] JESEME does not provide direct access to underlying texts to avoid copyright issues, but provides links to corpus providers and their search interfaces.

5.3.1. History of Electricity

Our first example is from the history of science and concerned with electricity. The modern scientific understanding of electrictiy can be traced back to the late 17$^{\text{th}}$ century, with rapid advances leading to technical applications in the 19$^{\text{th}}$ century. Whereas early researchers were limited to observing natural (weather) phenomena, electricity became part of controlled laboratory experiments during this time frame (Home, 2008, pp. 368–371)—its context and usage in both science and culture can thus be assumed to have changed strongly (Bertucci, 2007).
We used JeSemE to query both RSC and DTA for words related to electricity, i.e., *Elektrizität* ['electricity'], *electricity*, *elektrisch* ['electrical'], *electrical*, *Funken* ['sparks'] and *spark*. Our results matched scholarly expectations, suggesting a general applicability for studying other words of interest from a diachronic perspective.

Electricity became over time more and more similar to *spark*, *conductor* and *magnetism*, whereas its similarity to *lightning* decreased continuously, as shown in Figure 5.5. Its increased similarity to *magnetism* can be attributed to the discovery of electromagnetism in the early 19$^{\text{th}}$ century (Saslow, 2002, pp. 505–512). The discovery of electromagnetism also appears in DTA via an increase in the similarity of *Elektrizität* ['electricity'] and *Magnet* ['magnet'] during the middle of the 19$^{\text{th}}$ century. Both DTA and RSC reflect the decrease in the similarity of *electricity* to naturally occurring *lightning* (German: *Blitz*) over time, with a simultaneous increase of its similarity to *conductor* (German: *Leiter*). This coincides with the known shift of the notion of *electricity* away from natural phenomena and towards artificial creation and industrial applications (Morus, 2011).
The most strongly associated words (according to normalized χ^2, see Figure 5.6) for *electricity* (respectively German *Elektrizität*) are related to electrical charge, e.g., the adjective *negative*. RSC provides further indirect references to electrical charge through adjectives describing materials used in electric experiments, i.e., *vitreous* and *resinous*. Historically, these predate *negative* as well as *positive* and could be used interchangeably (Saslow, 2002, pp. 44–45). DTA also shows these associated materials, i.e., *Glas* ['glas'] and *Harz* ['resin'] (which again could be used to indicate charge,

see e.g., Lichtenberg & Erxleben (1787, p. 434)), but only from 1811 on. Association measures were overall less helpful for uncovering changes in historical understanding. The only interesting shift is the increased assocation with *magnetism*, however this change is less dynamic than those affecting words indicating charge.

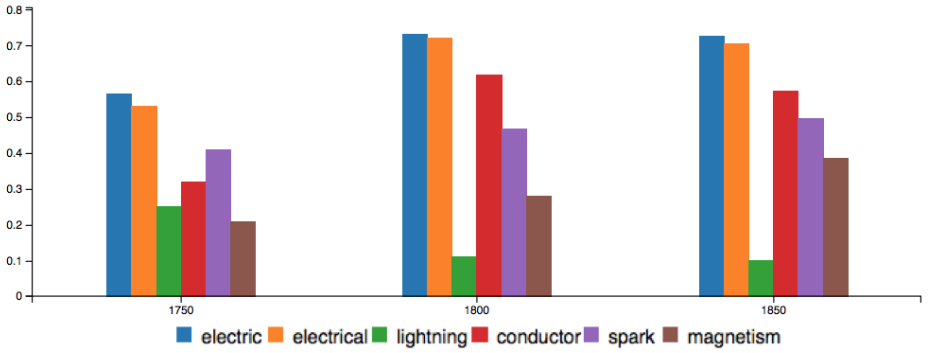

Figure 5.5: Similarity of *electricity* to selected reference words (y-axis, by cosine) in RSC; JeSemE screenshot with magnified legend.

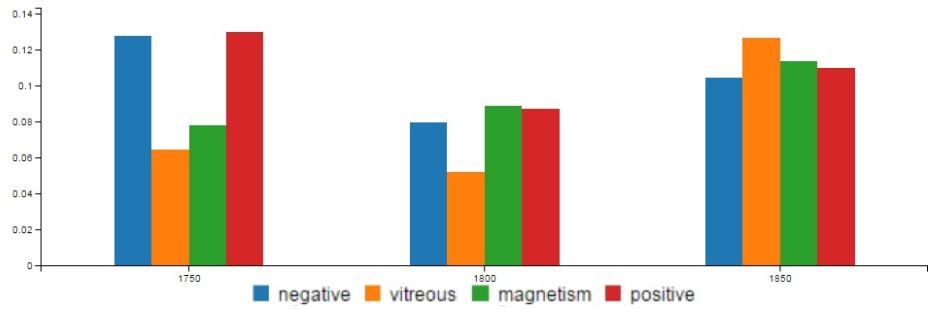

Figure 5.6: Association between *electricity* and selected reference words (y-axis, by normalized χ^2) in RSC; JeSemE screenshot with magnified legend.

electrical and its German counterpart *elektrisch* provide even better indicators for the increased understanding of the underlying phenomena in both RSC and DTA. Both SVD_{PPMI} and normalized χ^2 indicate *matter* and *spark* (*Materie* and *Funken*, respectively, in German) to be relevant, pointing at a connection between *electricity* and philosophical concepts. Especially in the late 18^{th} century, scientists assumed the existence of an 'electrical matter', an abstract and only vaguely defined force with possible links to life itself (Steigerwald, 2013). The idea of electricity as a matter was abandoned in the 19^{th} century, as can be seen by the steep decline of the corresponding words in terms of similarity in Figure 5.7; analogous observations can be made with regards to context specificity in DTA as well. The association between *electrical* and *spark* is relatively stable over time, whereas the association between their German counterparts, i.e., *elektrisch* and *Funken*, declines.

spark and its German counterpart *Funken* develop quite differently. As shown in Figure 5.8, the former becomes less and less similar to *fire*, indicating a semantic narrowing towards research on electricity. In contrast, the similarity of German *Funken* to *Flamme* ['flame'] and *Feuer* ['fire'] is rather constant. Sparks were a part of popular public experiments, e.g., by making the hair of someone glow, alluring to halos from Christian iconography (Hochadel, 2006, p. 528). Differences between DTA and RSC might be language specific, corpus specific (as DTA is not genre specific) or even due to different foci in research—German scientist were more eager to focus on electromagnetism (Morus, 1998).

We could also observe an interesting shift in word emotion for *Elektrizität* and *elektrisch* which rise in Dominance by about 1 standard deviation (maximal value of 0.65 for 1811–1840). In contrast, RSC always shows *electricity* and *electrical* to be average in Dominance, but very high in Arousal (up to 3.9 for *electrical*!). This is probably both an indicator for past scientists' excitement about a novel area of research as well as their scientific writing style.

We deem our investigation in the historical understanding of the notion of electricity by applying JeSemE to be fruitful, as our results match known developments. History of science is a relatively small field of study operating on large corpora and could thus greatly profit

from applying automatic methods to increase the efficiency of its research procedures and provide guidance for focusing on particularly relevant documents. The agreement between JeSemE's results and scholarly knowledge can be seen as preliminary evidence for the validity of our quantitative approach and spurred us to apply it to questions from another domain in the next section.

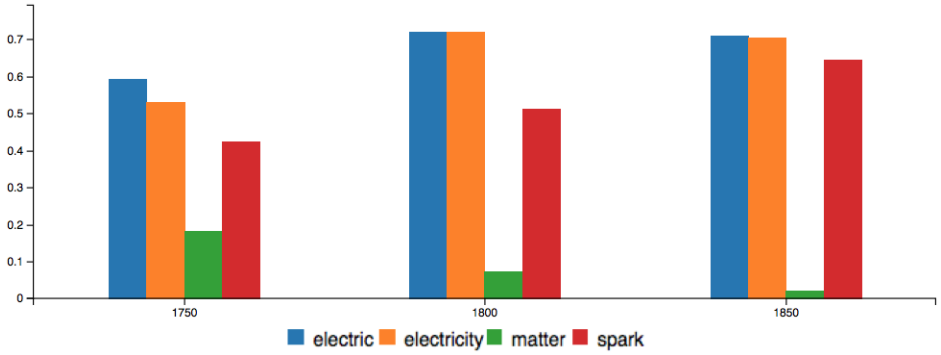

Figure 5.7: Similarity of *electrical* to selected reference words (y-axis, by cosine) in RSC; JeSemE screenshot with magnified legend.

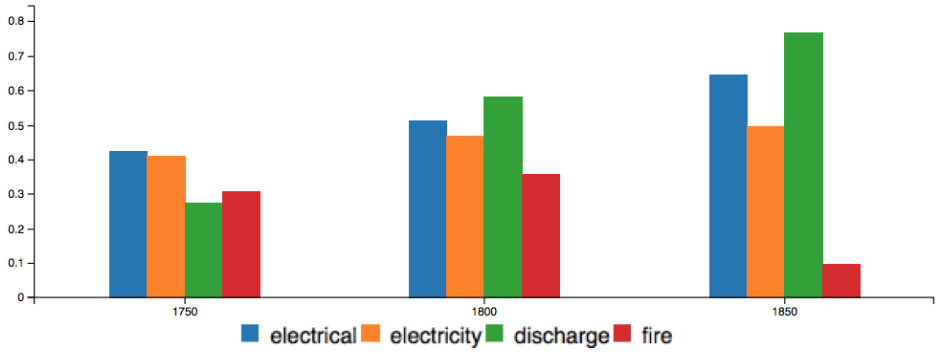

Figure 5.8: Similarity of *spark* to selected reference words (y-axis, by cosine) in RSC; JeSemE screenshot with magnified legend.

5.3.2. Words Linked to Romanticism

Our second example comes from literary studies and investigates how the meaning of *romantic* and other words linked with Romanticism changes over time. Romanticism can be roughly described as a artistic and especially literary movement in late 18th and 19th century Europe. It is commonly characterized by an increased interest in nature, the Middle Ages and individualism (Kremer & Kilcher, 2015; Perry, 1998; Stevens, 2004; Ziolkowski, 1969). This definition is highly simplified, as Romanticism was, e.g., also linked to political and, at least in Germany, scientific movements (Berlin, 2013; Jahn, 1994; Knight, 1990). However, it is sufficient to provide a temporal and lexical starting point and show how JeSemE could be used by scholars in the humanities.

We identified words linked to Romanticism by selecting the ten most frequent nouns in 'Des Knaben Wunderhorn', a romantic collection of short texts by von Arnim & Brentano (1806–1808) by using the annotations provided in DTA (see also Hellrich & Hahn (2016c)), i.e., *Gott* ['god'], *Herr* ['lord', 'Mr.'], *Liebe* ['love'], *Tag* ['day'], *Frau* ['woman', 'Ms.', 'wife'], *Mutter* ['mother'], *Herz* ['heart'], *Wein* ['wine'], *Nacht* ['night'] and *Mann* ['man', but not 'mankind']. Obviously, we were also interested in the word *Romantik* ['Romanticism'] itself as well as the adjective *romantisch* ['romantic'].

We queried JeSemE for these 12 German words (in GBG and DTA) and their English translations (in GBF and COHA). Due to JeSemE's minimum corpus size requirements and the resulting temporal coverage (see Section 5.2.1), we were mostly limited to modeling word meaning after Romanticism.[35]

[35] The temporal extension of Romanticism is not clearly agreed on, especially if multiple countries or non-literary arts are to be included. The longest extension in the aforementioned sources is given by Stevens (2004) as 1750 to 1850, the shortest by Kremer & Kilcher (2015) with about 1796 to about 1830. DTA is the only corpus in JeSemE, besides the domain specific and thus ill-suited RSC, to cover this time span. In contrast, GBF covers only the 1820s and later, whereas COHA and GBG begin with the 1830s (see Table 5.5).

Gott ['god'] rises[36] in similarity to *Jesus* in GBG (+0.5) as well as GBF and COHA (about +0.25 in both), whereas their similarity is rather volatile in DTA (drop from 0.44 to 0.03 for 1811–1840). Neither association values nor manually inspecting text samples provided an explanation for this development.
Differences in word emotions seem to follow a well known trend towards secularization (Chadwick, 1975), i.e., Valence falls in all corpora except DTA, but both initial and current Valence values are well above average. Arousal is about average in all corpora. Dominance is above average in the English corpora and seems to develop parallel to Valence, whereas it is below average in the German corpora. Figure 5.9 shows an increase in Valence and Dominance during the 1950s, 1970s and 2000s decades in COHA, possibly linked to the rise of Evangelicalism since the 1970s (Brinkley, 2003, p. 900) and the civil rights movement's adoption of religious language (Lippy, 2010, p. 253).

Herr ['lord' or 'Mr.'] seems to be increasingly used as a honorific in German, as indicated by its rising similarity to *geehrte* ['honored'] and *verehren* ['to honor']. Especially the former is used in letters as a German counterpart to *Dear...*. No relevant similarity changes could be observed in DTA. In English the two senses of *Herr* are signified by separate words, *lord* being the more interesting one. We found it to become more similar to *god* and less similar to *earl* over time, especially so in COHA. Emotion values in both German corpora resemble those for *lord*, especially in regards to high Valence.

Liebe ['love'] shows a marked trend in word emotions, but not in its most similar words. Both Valence and Dominance drop sharply—in GBG by nearly 2 standard deviations—in all corpora since the early 19$^{\text{th}}$ century (in DTA only after 1870). *Ehe* ['marriage'], which we deemed an interesting reference and added manually, is not rated as very similar (mostly below 0.1) in GBF, but shows a higher and increasing similarity in GBG (0.16–0.37) and COHA (0.13–0.28) and a constant low similarity (≈ 0.15) in DTA.

[36] Changes occur from the first to the last time span covered by a corpus, unless specified otherwise.

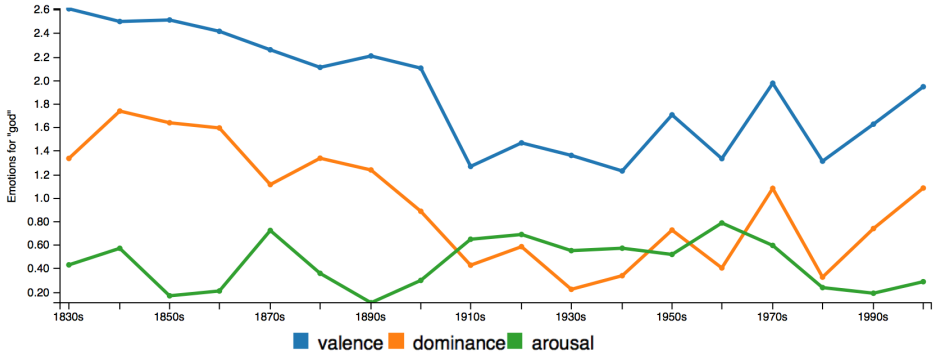

Figure 5.9: Emotional development of *god* in COHA (y-axis, in standard deviation from average); JESEMEscreenshot with magnified legend.

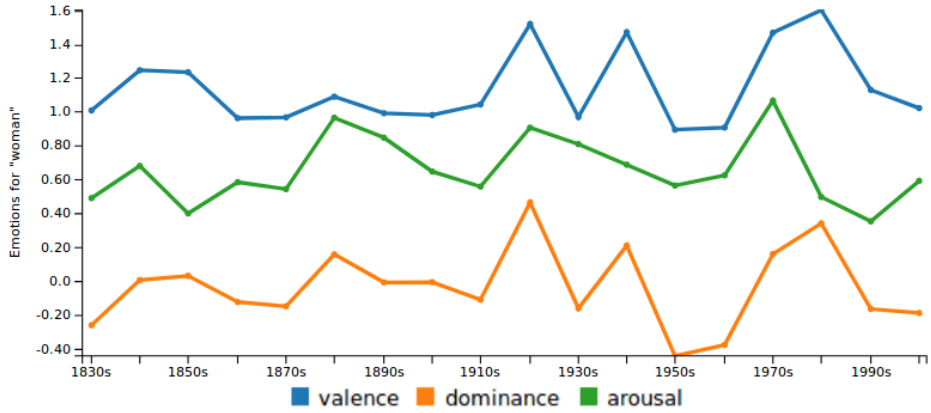

Figure 5.10: Emotional development of *woman* in COHA (y-axis, in standard deviation from average); JESEMEscreenshot with magnified legend.

Tag & Nacht ['day' & 'night'] show a surprisingly low mutual similarity in German (~ 0.35) compared to English (~ 0.6) corpora. We could identify two patterns typical for *Nacht*, but not *Tag*, i.e., *Nacht vom _ zum _* and *Nacht vom _ auf _* [both 'night from _ to _']. We assume these frequent patterns—e.g., for 1900 there are 924 5-grams starting with *Nacht* in GBG, of which 281 use the *zum*-pattern and 255 the *auf*-pattern—to be the most likely explanation for the low similarity values in German.

Frau ['woman', 'Ms.' or 'wife'] is rather similar to family related words, i.e., *Ehe* ['marriage'], *Kind* ['child'], *Mutter* ['mother'] and *Familie* ['family'] in the German corpora. This is less pronounced for *woman* in the English corpora for which *man* and *girl* are most similar. The GBG shows a long term upwards trend (+0.53) in similarity with *Familie*, but also a drop in similarity for all family related words in the 1980s in GBG, both could be caused by changes in societal expectations.

Both English corpora show very static most similar words, but dynamic word emotions. The latter effect is especially clear in COHA, as shown in Figure 5.10. Spikes in Valence and Dominance during the 1920s, 1940s, 1970s and 1980s match turning points in US women's history, i.e., first-wave feminism leading to women's suffrage in 1920, increased female workplace participation during WW2 as well as second-wave feminism leading to better education, legalized abortion and greater political equality since about 1970 (Brinkley, 2003, pp. 586–587,760,871–874). Earlier smaller spikes might be connected to first feminist movements in the 1830s and 1840s and the growth of feminist organizations (Brinkley, 2003, pp. 333–334). In contrast, emotion values in GBG tend towards average, with the exception of a short counter movement in the 1940s. This might be linked to the NS regime's propaganda[37] and glorification of motherhood (Frietsch & Herkommer, 2015), but is more pronounced here than for *Mutter* ['mother'] (see next example).

[37] However, Koplenig (2017) showed a rise in Swiss German documents during this time, which should counteract NS propaganda.

Mutter ['mother'] is always very similar to words for other family members, e.g., *father*, *husband* or *sister* (respectively their German equivalents). All corpora show a long term drop in Valence (e.g., −1.2 in GBG). In contrast, Arousal drops only in the English corpora, but rises in GBG. COHA (but not GBF) shows spikes in Valence for the 1940s and 1980s, which also appear in GBG to a lesser degree (together with a spike in the 1880s). According to Brinkley (2003, p. 658) motherhood did become less linked to instinctual behavior during the 1920s, however we could find no matching changes with JeSemE. No overall changes could be observed in DTA.

Herz ['heart'] can be used anatomically, metaphorically or metonymically (see e.g., Grimm & Grimm (1999a, cols. 1207–1223), Simpson & Weiner (1989a, pp. 60–65)). Both metaphorical and metonymical usage were historically common despite the long-known anatomical function (Aird, 2011; Niemeier, 2003). Figure 5.4 (previously used to illustrate JeSemE's `result` page) shows its increasing similarity with *lungs* and *stroke* and decreasing similarity with *soul* in COHA, pointing towards a more anatomical usage. Both GBF and GBG also show an increased similarity with other organs (e.g., *Magen* ['stomach']), but no drop in similarity with *soul* (respectively German *Seele*). No lasting changes could be observed in DTA.

Word emotions in COHA, GBF and GBG show a clear drop (e.g., ≈ 1.2 in COHA) in both Valence and Dominance. This might be due to our ability to 'change our heart' in a metaphorical sense, while we have little control over our anatomical heart and can be threatened by cardiovascular diseases (such as *stroke*). This emotional change seems to predate the similarity changes in both English corpora but not in GBG.

Interestingly, COHA also shows *hearts* to become less and less similar to *heart*, as it seems to be more consistently used metaphorically, e.g., *and their foolish hearts were darkened*.[38]

[38] According to COHA's online version (avoids blackened passage; no direct links possible) an excerpt from LaHaye & Jenkins (2007), provided without page number.

Wein ['wine'] is overall static in its similarity values and mostly also in its emotion values. GBG shows an increase in similarity with both foods and drinks, e.g., +0.24 for *Milch* 'milk'. COHA shows both a long term increase in Valence and Dominance as well as rapid movement (≈ 1 standard deviation) between the 1910s and 1930s. We assume a link with (alcohol) prohibition in the USA (1920–1933; Brinkley (2003, p. 665)), as *drink* is similarly effected; other alcoholic beverages are not covered.

Mann ['man', but not 'mankind'] is static in its most similar words in GBF, COHA and DTA. However, it shows two puzzling most similar words in GBG, i.e., *Besatzung* ['crew', 'occupation'] and *5000*, both becoming dissimilar during the 20$^{\text{th}}$ century. *Besatzung* seems to be increasingly used in the sense of 'occupation' instead of 'crew', which would explain less connection with *Mann*. Its word emotions also shift strongly around 1900 (Arousal rises while Valence and Dominance drop by about 2 standard deviations) and its association with *unter* ['under'] rises, due to 5-grams such as *unter der deutschen Besatzung gelitten* ['suffered under German occupation'] (from GBG for 2000). No such explanation could be found for *5000* which is mainly similar to (and associated with) other numbers.

Romantik & romantisch ['Romanticism', 'romantic'] were harder to study than other words, as *Romantik* is only covered in GBG, but neither DTA nor the two English corpora did contain it (respectively *Romanticism*). We will thus discuss *Romantik*, *romantisch* and *romantic* together.

As shown in Figure 5.11, *Romantik* is initially most similar to *Griechische* ['Greek'] and *Protestantismus* ['Protestantism'], a connection we cannot explain. However, other similarity trends seem to be linked with literary scholarship, e.g., an increase in similarity to *Realismus* ['Realism'] and *klassisch* ['classical'].

Author names also show the influence of literary scholarship. Similarity to *Schlegel*[39] is high in the 1850s and 1860s, drops in the 1870s and rises to a relatively stable value during the late 19$^{\text{th}}$ century.

[39] The brothers August Wilhelm and Friedrich Schlegel (Anz, 2010; Schlaffer, 2010b) were very influential during Romanticism. The earlier Johann Elias Schlegel (Hollmer, 2010) is not connected with Romanticism.

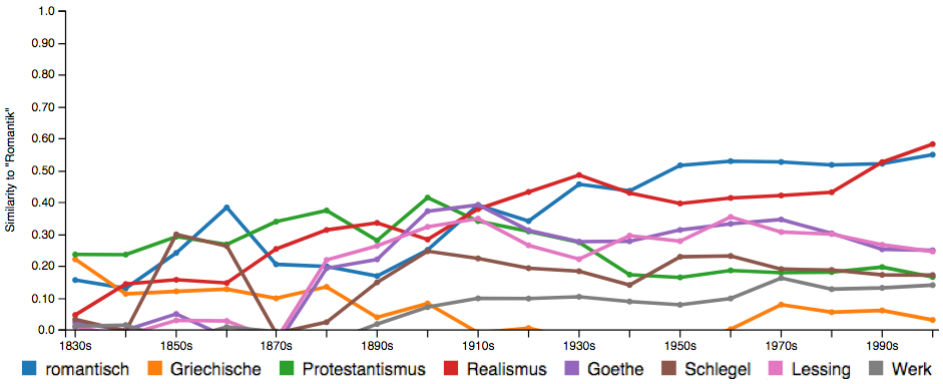

Figure 5.11: Similarity of *Romantik* ['Romanticism'] to selected reference words (y-axis, by cosine) in GBG; JeSemE screenshot with magnified legend.

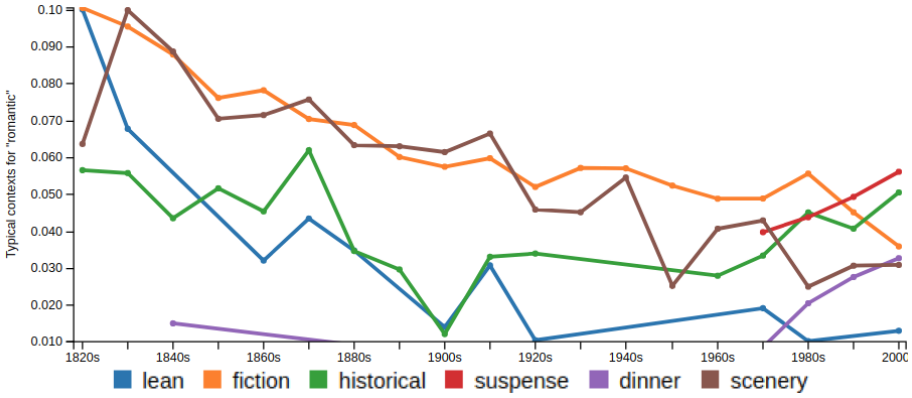

Figure 5.12: Association of *romantic* with selected reference words (y-axis, by normalized χ^2) in GBF; JeSemE screenshot with magnified legend.

While the *Schlegel* brothers were highly involved in Romanticism, the names of other authors show a similar trend since the late 19th century, e.g., *Goethe*[40] and *Lessing*.[41]
There are two non-exclusive explanations for the influence of literary scholarship on the meaning of *Romantik*. First, the composition of GBG shifted towards academic texts—recall the rise of brackets shown in Figure 3.5. Second, Romanticism is claimed to be increasingly studied (Kremer & Kilcher, 2015, pp. 52–53). Both explanations would also fit *Romantik*'s trend towards average emotions since the late 19th century and its increasing frequency (the latter only between the 1890s and 1930s).

The adjective *romantisch*, respectively its English counterpart *romantic*,[42] could be modeled in all corpora. Both originally referred to something appearing fiction-like[43] and were historically also used in landscape descriptions, especially during the 18th century (see Grimm & Grimm (1999b, cols. 1155–1157), Simpson & Weiner (1989b, pp. 65–66)).
Results for GBG also show the influence of literary studies' texts. Similarity to *klassisch* ['classical'] is consistently high and similarity to *Realismus* ['realism'] rises, especially from the 1900s on. Author names (i.e., *Goethe*, *Lessing*, *Schlegel*) also rise in similarity, but slightly later and less sudden. Similarity to *Ironie* ['irony'], which is seen as an important part or romantic poetic theory (Kremer & Kilcher, 2015, pp. 92–95), rises after the 1950s. Word emotions do again trend towards average after an increase in the middle of the 19th century.

DTA appears to be unaffected by literary studies' texts. *Lessing* is the only modeled author name[44] and not very similar ($\leq$ 0.12). Similarity to other words is very low in general, with the initially most similar *anmutig* ['graceful'] being replaced by *jugendlich* ['youthful'] and *Jugend* ['youth'] towards 1900. Association with *Gegend* ['area'] drops, indicating *romantisch* to be no

[40] Johann Wolfgang von *Goethe*'s affiliation with Romanticism is disputed, see e.g., Ashton (1998, p. 496), Kremer & Kilcher (2015, p. 1) or Schlaffer (2010a)).
[41] Gotthold Ephraim *Lessing* was an Enlightenment author (Vollhardt, 2010).
[42] Our analysis does not distinguish between adjective and noun.
[43] They are derived from French *romant* ['novel'] (see Grimm & Grimm (1999b, cols. 1155–1157), Simpson & Weiner (1989b, pp. 65–66)).
[44] Other authors only rose to popularity after the initial 1751–1780 time span.

longer used in landscape descriptions. Valence and Dominance rise by more than 1.5 standard deviations over an initially average value, roughly matching trends for the 19th century in GBG.
GBF shows a rise in similarity with *suspense* (since the 1950s), love (since the 1970s) and *romance* (especially from the 1990s on). Association (see Figure 5.12) with *scenery* and *fiction* drops gradually. In contrast, *dinner* becomes more and more associated from the 1980s on. The initially high association with *lean* is due to the place name *Lough Lean*.[45] All three emotion dimensions rise in the late 20th century and are constantly above average. Overall, results suggest a historic popularity of *romantic* in landscape descriptions, with more recent trends probably pointing towards genre literature or courtship.

Finally, in COHA *romantic*'s similarity with *romance* oscillates around 0.3, whereas *incident*[46] becomes less similar while *intimate* becomes more similar during the first half of the 20th century. Association with *dinner* and *hopeless* increases during the 20th century. Word emotions are roughly similar to those in GBF, but more dynamic. All dimensions rise towards the 2000s and are constantly above average. Arousal is most dynamic, however changes do not appear to match historical events. COHA seems to show a change away from *romantic* describing events that are or appear fictive towards *romantic* describing courtship (*intimate*).

Overall, we find JeSemE's results for the tested words related to Romanticism to be insightful, but probably more relevant for historians than for literary scholars—most changes we observed can be linked with societal or historical developments, e.g., the women's rights movement. Few are implausible, e.g., the low similarity between *Gott* and *Jesus* in DTA for 1811–1840. Especially the application to *Romantik/romantisch/romantic* showed a strong corpus dependence, each highlighting a different aspect of Romanticism's influence. While

[45] See the following 1826 5-gram: *surrounds the romantic Lough Lean.* There is no entry in the place name dictionary by Mills (2011), but a Lough Leane can be found in Ireland [Accessed May 28th 2019]: `https://www.google.de/maps/@52.0377986,-9.5664993,14.25z`

[46] Used to describe exploration, as in the following text passage from COHA's online version (attributed to O'Brien (1921) without further details): *The Mutiny of the Bounty, perhaps the most romantic incident [...].*

GBG was influenced by literary studies, DTA and both English corpora seem to show genuine semantic change away from landscape descriptions and towards courtship.

5.4. Discussion

Our distributional approach towards semantic change seems to be well suited for further studies. We could detect several plausible changes in emotional connotation with JeSemE, e.g., the lowered Valence for *Gott/god* coincident with a trend towards secularization. We could also detect changes in denotation for *romantisch*, *romantic* and arguably also *Herz*, *lord* and *heart*—the first two acquired a new metaphorical sense, whereas the latter three were less often used in an existing metaphorical sense.
Our case studies would not have been possible with any of the other systems described in Section 5.2.4, as they lack in corpus coverage, information on historical word emotions, and often also word similarity measurements. JeSemE's unique capability to track emotional connotations should be interesting for many scholars, e.g., literary scholars interested in word change affecting later reception—recently widely discussed in regards to antiquated language and racism in children's books (Hahn et al., 2015). It could also be useful for social scientists or historians, as demonstrated by emotional changes for *woman* or *wine* in COHA being aligned with historical events.
Regarding corpora, GBG seems to be ill-suited for diachronic studies, due to changes in its composition which lead to artifacts, e.g., the similarity of different author names to *Romantik*. DTA showed few changes in Section 5.3.2, but seemed to be well suited for studying the history of science in Section 5.3.1. The domain specific RSC also proved to be well-suited for studying the history of science. For general language change GBF and COHA often provided similar results.
We could not identify any delays between historic events and measurable effects in corpora, in contrast to delays of about one decade in studies by Bentley et al. (2014) and Tahmasebi & Risse (2017a). Most of the trends we observed are probably too long-term for such an effect to be discernible. A notable exception was COHA showing changes connected with distinct historical events, e.g., Prohibition,

without any delays, probably due to the inclusion of magazines and newspapers.

While the results of our case studies are plausible, they can only provide limited proof of JeSemE's validity. Additional case studies or the creation of a semantic change gold standard (see Section 2.5.4) would be necessary to quantify its validity.

Chapter 6

Conclusion

Word embeddings can be used as a tool for diachronic linguistics and the digital humanities. Chapter 2 provided background information on relevant algorithms, whereas Chapter 3 introduced relevant diachronic corpora. The reliability of word embedding algorithms was studied in Chapter 4. Finally, Chapter 5 introduced a novel way to model word emotions and presented the JeSemE website as well as two case studies.

Reliability issues affect many popular embedding algorithms, i.e., they produce different word embeddings (and thus judgments on word similarity) when experiments are repeated. Such differences can be highly misleading in studies using the most similar words to track and visualize semantic change. This, in turn, makes many studies hard to reproduce,[1] diminishing their scientific value. To the best of my knowledge, I was the first to study the (lacking) reliability of word embeddings. Luckily, I found variants of the SVD_{PPMI} algorithm—especially my novel $\text{SVD}_{\text{wPPMI}}$ (see Section 4.6)—to be perfectly reliable. This problem is not specific to diachronic research, but can also affect synchronic studies on, e.g., language variation (Kulkarni et al., 2016), gender roles (Bolukbasi et al., 2016a) or social media content (see e.g., Preoţiuc-Pietro et al. (2016), Arendt & Volkova (2017)).

Emotional connotation information was derived using a bootstrapping process developed together with Sven Buechel. The process uses

[1] Following Ivie & Thain (2018), reproduction consists in carrying "out tasks that are equivalent in substance to the original, but may differ in ways that are not expected to be significant to the final result" (Ivie & Thain, 2018, p. 63:4).

word similarity judgments derived from historical texts to project current fine-grained emotional information on historical language. We evaluated this method by creating the first gold standard for historical word emotions based on expert judgments.
The Jena Semantic Explorer (JeSemE) website gives non-technical users access to state-of-the-art distributional semantics. It tracks changes in both denotation and emotional connotation as well as word association and frequency. It provides access to five diachronic corpora, i.e., the Corpus of Historical American English, the Deutsches Textarchiv Kernkorpus, the German and English Fiction sub-corpora of the Google Books Ngram corpus and the Royal Society Corpus.
The digital humanities case studies used JeSemE to investigate the history of science as well as words relevant to Romanticism. They showed JeSemE to produce plausible results and thus indicate that such a distributional approach is suitable.
Overall, my research provides methodological insights relevant for general applications of word embeddings as well as tailored solutions for diachronic research. Scholars interested in the semantic change of German or English during the last two centuries can now use JeSemE to easily utilize reliable state-of-the-art methods.
There are several interesting areas for future research:

- Additional studies on the unreliable nature of word embeddings might provide insight into the structure of the space described by them. Embedding spaces were already shown to behave in peculiar ways, e.g., word embeddings populate only a small part of the available space (Mimno & Thompson, 2017). While word embeddings are unquestionably well-suited for judging the similarity of words and compute analogies, it still needs to be investigated how and why they can achieve this (Arora et al., 2016; Gittens et al., 2017; Patel & Bhattacharyya, 2017). The reliability of contextualized word embeddings (e.g., BERT (Devlin et al., 2018) or ELMo (Peters et al., 2018)) should also be investigated.

- It seems advisable to investigate the reliability of other popular digital humanities methods, e.g., Topic modeling via Latent Dirichlet Allocation (Blei et al., 2003; Jockers, 2013; Mimno, 2012; Schöch, 2017). Quantifying topic modeling's (unreliable)

probabilistic nature[2] might lead to a resurgence of SVD-based methods (Deerwester et al., 1990) which are reliable yet need performance improvements.[3]

- Alternative approaches for creating low-memory word (vector) representations might be worth exploring, e.g., PPMI vectors with a minimum association (Levy et al., 2015) or a maximum number of associated contexts per modeled word (Gamallo, 2017). While such representations currently perform worse than word embeddings, they are reliable and at least some seem to be immune to artifacts caused by dimensionality reduction (Dubossarsky et al., 2017).

- Modeling historical emotional connotation is still a fledgling field and might profit from ongoing research on modeling emotions in general, e.g., using deep learning (Buechel & Hahn, 2018b). It would also be helpful to have larger gold standards for evaluation.

- Diachronic semantic research in general is hampered by this lack of gold standards, often limiting evaluation to qualitative assessments of plausibility (see Section 2.5.4).

- Finally, JeSemE was merely used in two case studies. Others could adopt JeSemE for their studies, e.g., studies in the history of science. This might, however, necessitate the addition of further corpora from other domains or languages.

[2] For example, we found only about half of the resulting topics to be stable during repeated experiments in an unrelated study (Hellrich & Rzymski, 2019).

[3] Their inability to add new documents to a collection without reprocessing the whole collection should be of little relevance for applications in the digital humanities.

Bibliography

Entries for conference proceedings were harmonized (e.g., order of date and location information) and prefixed with commonly used abbreviations.

Alberto Acerbi, Vasileios Lampos, Philip Garnett & R. Alexander Bentley (2013): The expression of emotions in 20th century books. In: *PLoS ONE*, 8(3): e59030.

William C. Aird (2011): Discovery of the cardiovascular system: from Galen to William Harvey. In: *Journal of Thrombosis and Haemostasis*, 9(s1): 118–129.

Rami Al-Rfou, Bryan Perozzi & Steven Skiena (2013): Polyglot: Distributed Word Representations for Multilingual NLP. In: *CoNLL-2013 — Proceedings of the Seventeenth Conference on Computational Natural Language Learning. Sofia, Bulgaria, August 8–9, 2013*, pp. 183–192.

Maria Antoniak & David Mimno (2018): Evaluating the Stability of Embedding-based Word Similarities. In: *Transactions of the Association for Computational Linguistics*, 6: 107–120.

Thomas Anz (2010): Schlegel, Friedrich. In: Bernd Lutz & Benedikt Jeßing (eds.) *Metzler Autorenlexikon. Deutschsprachige Dichter und Schriftsteller vom Mittelalter bis zur Gegenwart*, pp. 681–682. J.B. Metzler, 4th edition.

Dustin Arendt & Svitlana Volkova (2017): ESTEEM: A Novel Framework for Qualitatively Evaluating and Visualizing Spatiotemporal Embeddings in Social Media. In: *ACL 2017 — Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Vancouver, Canada, July 30 – August 4, 2017*, pp. 25–30.

Achim von Arnim & Clemens Brentano (1806–1808): *Des Knaben Wunderhorn*, volume 1–3. Mohr und Zimmer. Annotated TCF version provided by the Deutsches Textarchiv.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma & Andrej Risteski (2016): A Latent Variable Model Approach to PMI-based Word Embeddings. In: *Transactions of the Association of Computational Linguistics*, 4: 385–399.

Rosemary Ashton (1998): England and Germany. In: Duncan Wu (ed.) *A Companion to Romanticism*, pp. 495–504. Blackwell.

Robert Bamler & Stephan Mandt (2017): Dynamic Word Embeddings. In: *ICML 2017 — Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia, August 6–11, 2017*, pp. 380–389.

Marco Baroni, Georgiana Dinu & Germán Kruszewski (2014): *Don't count, predict!* A systematic comparison of context-counting vs. context-predicting semantic vectors. In: *ACL 2014 — Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Long Papers. Baltimore, MD, USA, June 22–27, 2014*, pp. 238–247.

Marco Baroni & Alessandro Lenci (2010): Distributional Memory: A General Framework for Corpus-Based Semantics. In: *Computational Linguistics*, 36(4): 673–721.

Miroslav Batchkarov, Thomas Kober, Jeremy Reffin, Julie Weeds & David Weir (2016): A critique of word similarity as a method for evaluating distributional semantic models. In: *RepEval 2016 — Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP @ ACL 2016. Berlin, Germany, August 12, 2016*, pp. 7–12.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent & Christian Jauvin (2003): A Neural Probabilistic Language Model. In: *Journal Of Machine Learning Research*, 3: 1137–1155.

R. Alexander Bentley, Alberto Acerbi, Paul Ormerod & Vasileios Lampos (2014): Books average previous decade of economic misery. In: *PLoS ONE*, 9(1): e83147.

Isaiah Berlin (2013): *The Roots of Romanticism.* Princeton University Press, 2nd edition.

M. W. Berry (1992): Large-scale sparse singular value computations. In: *The International Journal of Supercomputer Applications*, 6(1): 13–49.

Paola Bertucci (2007): Sparks in the dark: the attraction of electricity in the eighteenth century. In: *Endeavour*, 31(3): 88–93.

Yves Bestgen (2008): Building affective lexicons from specific corpora for automatic sentiment analysis. In: *LREC 2008 — Proceedings of the Sixth International Conference on Language Resources and Evaluation. Marrakech, Morocco, May 28–30, 2008*, pp. 496–500.

Yves Bestgen & Nadja Vincze (2012): Checking and bootstrapping lexical norms by means of word similarity indexes. In: *Behavior Research Methods*, 44(4): 998–1006.

Douglas Biber, Susan Conrad & Randi Reppen (2000): *Corpus Linguistics — Investigating language structure and use*. Cambridge University Press.

Chris Biemann (2006): Chinese Whispers - an Efficient Graph Clustering Algorithm and its Application to Natural Language Processing Problems. In: *TextGraphs 2006 — Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing @ HLT-NAACL 2006. New York City, NY, USA, June 9, 2006*, pp. 73–80.

Andreas Blank (1999): Why do new meanings occur? A cognitive typology of the motivations for lexical Semantic change. In: Andreas Blank & Peter Koch (eds.) *Historical Semantics and Cognition*, pp. 61–89. De Gruyter Mouton.

David M. Blei, Andrew Y. Ng & Michael I. Jordan (2003): Latent Dirichlet Allocation. In: *The Journal of Machine Learning Research*, 3: 993–1022.

Leonard Bloomfield (1984): *Language*. University of Chicago Press. [Reprint, first published 1933].

Piotr Bojanowski, Edouard Grave, Armand Joulin & Tomas Mikolov (2017): Enriching Word Vectors with Subword Information. In: *Transactions of the Association of Computational Linguistics*, 5: 135–146.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama & Adam Kalai (2016a): Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In: *NIPS 2016 — Advances in Neural Information Processing Systems 29. Barcelona, Spain, December 5–10, 2016*, pp. 4349–4357.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama & Adam Kalai (2016b): Quantifying and reducing stereotypes in word embeddings. In: *Proceedings of the Workshop on #Data4Good:*

Machine Learning in Social Good Applications @ ICML 2016. New York City, NY, USA, June 24, 2016, pp. 41–45. https://arxiv.org/pdf/1606.06121.pdf [Accessed May 28th 2019].

Stefan Bordag (2008): A Comparison of Co-occurrence and Similarity Measures as Simulations of Context. In: *CICLing 2008 — Computational Linguistics and Intelligent Text Processing: 9th International Conference. Haifa, Israel, February 17–23, 2008*, pp. 52–63.

Harold Borko & Myrna Bernick (1963): Automatic Document Classification. In: *Journal of the ACM*, 10(2): 151–162.

Margaret M. Bradley & Peter J. Lang (1994): Measuring emotion: The self-assessment manikin and the semantic differential. In: *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1): 49–59.

Margaret M. Bradley & Peter J. Lang (1999): *Affective Norms for English Words (*ANEW*): Stimuli, Instruction Manual and Affective Ratings.* Technical Report C-1, The Center for Research in Psychophysiology, University of Florida. https://www.uvm.edu/pdodds/teaching/courses/2009-08UVM-300/docs/others/everything/bradley1999a.pdf [Accessed May 28th 2019].

Alan Brinkley (2003): *American History. A Survey.* McGraw Hill, 11th edition.

Elia Bruni, Gemma Boleda, Marco Baroni & Nam-Khanh Tran (2012): Distributional Semantics in Technicolor. In: *ACL 2012 — Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers. Jeju Island, Republic of Korea, July 8–14, 2012*, pp. 136–145.

Sven Buechel & Udo Hahn (2016): Emotion analysis as a regression problem: Dimensional models and their implications on emotion representation and metrical evaluation. In: *ECAI 2016 — Proceedings of the 22nd European Conference on Artificial Intelligence: Long Papers. The Hague, The Netherlands, August 29 – September 2, 2016*, pp. 1114–1122.

Sven Buechel & Udo Hahn (2018a): Representation Mapping: A Novel Approach to Generate High-Quality Multi-Lingual Emotion Lexicons. In: *LREC 2018 — Proceedings of the Eleventh International Conference on Language Resources and Evaluation. Miyazaki, Japan, May 7–12, 2018*, pp. 184–191.

Sven Buechel & Udo Hahn (2018b): Word Emotion Induction for Multiple Languages as a Deep Multi-Task Learning Problem. In: *NAACL-HLT 2018 — Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Long Papers. New Orleans, LA, USA, June 2–4, 2018*, pp. 1907–1918.

Sven Buechel, Johannes Hellrich & Udo Hahn (2016): Feelings from the Past: adapting Affective Lexicons for Historical Emotion Analysis. In: *LT4DH — Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities @ COLING 2016. December 11, 2016, Osaka, Japan*, pp. 54–61.

Sven Buechel, Johannes Hellrich & Udo Hahn (2017): The Course of Emotion in Three Centuries of German Text—A Methodical Framework. In: *Digital Humanities 2017 — Conference Abstracts of the 2017 Conference of the Alliance of Digital Humanities Organizations (ADHO). Montréal, Quebec, Canada, August 8–11, 2017*, pp. 176–179.

John A. Bullinaria & Joseph P. Levy (2007): Extracting semantic representations from word co-occurrence statistics: a computational study. In: *Behavior Research Methods*, 39(3): 510–526.

John A. Bullinaria & Joseph P. Levy (2012): Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. In: *Behavior Research Methods*, 44(3): 890–907.

Aylin Caliskan, Joanna J. Bryson & Arvind Narayanan (2017): Semantics derived automatically from language corpora contain human-like biases. In: *Science*, 356(6334): 183–186.

Rafael A. Calvo & Sunghwan Mac Kim (2013): Emotions in text: Dimensional and categorical models. In: *Computational Intelligence*, 29(3): 527–543.

Edward G. Carmines & Richard A. Zeller (1992): *Reliability and validity assessment*. Sage Publications, 17th edition.

Owen Chadwick (1975): *The secularization of the European Mind in the nineteenth century*. Cambridge University Press.

Mansi Chugh, Peter A. Whigham & Grant Dick (2018): Stability of Word Embeddings Using Word2Vec. In: Tanja Mitrovic, Bing Xue & Xiaodong Li (eds.) *Advances in Artificial Intelligence. AI 2018 — Proceedings of the 31st Australasian Joint Conference on Artificial Intelligence. Wellington, New Zealand. December 11–14, 2018*, pp. 812–818.

Kenneth Ward Church & Patrick Hanks (1990): Word Association Norms, Mutual Information, and Lexicography. In: *Computational Linguistics*, 16(1): 22–29.

Stephen Clark (2015): Vector Space Models of Lexical Meaning. In: Shalom Lappin & Chris Fox (eds.) *The Handbook of Contemporary Semantic Theory*, pp. 493–522. John Wiley & Sons.

N.E. Collinge (1990): Language as it evolves: tracing its forms and families. In: N.E. Collinge (ed.) *An encyclopedia of Language*, pp. 876–916. Routledge.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu & Pavel Kuksa (2011): Natural Language Processing (Almost) from Scratch. In: *Journal Of Machine Learning Research*, 12: 2493–2537.

Paul Cook & Suzanne Stevenson (2010): Automatically identifying changes in the semantic orientation of words. In: *LREC 2010 — Proceedings of the 7th International Conference on Language Resources and Evaluation. La Valletta, Malta, May 17–23, 2010*, pp. 28–34.

Anne Curzan (2009): Historical corpus linguistics and evidence of language change. In: Anke Lüdeling & Merja Kytö (eds.) *Corpus Linguistics. An International Handbook*, pp. 1091–1109. Mouton de Gruyter.

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

Mark Davies (2014): Making Google Books n-grams useful for a wide range of research on language change. In: *International Journal of Corpus Linguistics*, 19(3): 401–416.

Scott C. Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer & Richard Harshman (1990): Indexing by latent semantic analysis. In: *Journal of the American Society for Information Science*, 41(6): 391–407.

Fermín Moscoso del Prado Martín & Christian Brendel (2016): Case and Cause in Icelandic: Reconstructing Causal Networks of Cascaded Language Changes. In: *ACL 2016 — Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Long Papers. Berlin, Germany, August 7–12, 2016*, pp. 2421–2430.

Marco Del Tredici, Raquel Fernández & Gemma Boleda (2019):

Short-Term Meaning Shift: A Distributional Exploration. In: *NAACL-HLT 2019 — Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Long and Short Papers. Minneapolis, MN, USA, June 2–7, 2019*, pp. 2069–2075.

Janez Demšar (2006): Statistical comparisons of classifiers over multiple data sets. In: *Journal of Machine Learning Research*, 7: 1–30.

Jacob Devlin, Ming-Wei Chang, Kenton Lee & Kristina Toutanova (2018): BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. `http://arxiv.org/abs/1810.04805` [Accessed June 5th 2019].

Beate Dorow (2006): *A Graph Model for Words and their Meanings*. Ph.D. thesis, Universität Stuttgart. `http://dx.doi.org/10.18419/opus-2601` [Accessed May 28th 2019].

Lauren B. Doyle (1961): Semantic Road Maps for Literature Searchers. In: *Journal of the ACM*, 8(4): 553–578.

Haim Dubossarsky (2018): *Semantic change at large: A computational approach for semantic change research*. Ph.D. thesis, Hebrew University of Jerusalem. `http://www.cs.huji.ac.il/~daphna/theses/Haim_Dubossarsky_2018.pdf` [Accessed May 28th 2019].

Haim Dubossarsky, Eitan Grossman & Daphna Weinshall (2017): Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models. In: *EMNLP 2017 — Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, September 7–11, 2017*, pp. 1136–1145.

Haim Dubossarsky, Yulia Tsvetkov, Chris Dyer & Eitan Grossman (2015): A bottom up approach to category mapping and meaning change. In Word Structure and Word Usage. In: *NetWordS 2015 — Proceedings of the NetWordS Final Conference on Word Knowledge and Word Usage: Representations and Processes in the Mental Lexicon. Pisa, Italy, March 30 – April 1, 2015*, pp. 66–70.

Haim Dubossarsky, Daphna Weinshall & Eitan Grossman (2016): Verbs change more than nouns: A bottom up computational approach to semantic change. In: *Lingue e Linguaggio*, XV(1): 7–28.

Steffen Eger & Alexander Mehler (2016): On the Linearity of Semantic Change: Investigating Meaning Variation via Dynamic

Graph Models. In: *ACL 2016 — Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Short Papers. Berlin, Germany, August 7–12, 2016*, pp. 52–58.

Paul Ekman (1992): An argument for basic emotions. In: *Cognition & Emotion*, 6(3–4): 169–200.

Adrian Englhardt, Jens Willkomm, Martin Schäler & Klemens Böhm (2019): Improving semantic change analysis by combining word embeddings and word frequencies. In: *International Journal on Digital Libraries*, pp. 1–18. `https://doi.org/10.1007/s00799-019-00271-6` [Accessed May 28th 2019].

Stefan Evert (2005): *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, IMS, University of Stuttgart. `http://dx.doi.org/10.18419/opus-2556` [Accessed May 28th 2019].

Stefan Evert & Brigitte Krenn (2001): Methods for the Qualitative Evaluation of Lexical Association Measures. In: *ACL 2001 — Proceedings of 39th Annual Meeting of the Association for Computational Linguistics: Short Papers. July 9–11, 2001, Toulouse, France*, pp. 188–195.

R. Fano (1966): *Transmission of Information: A Statistical Theory of Communications*. MIT Press. [Reprint, first published 1961].

Murhaf Fares, Andrey Kutuzov, Stephan Oepen & Erik Velldal (2017): Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In: *NoDaLiDa 2017 — Proceedings of the 21st Nordic Conference on Computational Linguistics. Gothenburg, Sweden, May 22–24, 2017*, pp. 271–276.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy & Noah A. Smith (2015): Retrofitting Word Vectors to Semantic Lexicons. In: *NAACL-HLT 2015 — Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, CO, USA, May 31 – June 5, 2015*, pp. 1606–1615.

Christiane Fellbaum (ed.) (1998): WordNet*: An Electronic Lexical Database*. MIT Press.

Olivier Ferret (2017): Turning Distributional Thesauri into Word Vectors for Synonym Extraction and Expansion. In: *IJCNLP 2017 — Proceedings of the Eighth International Joint Conference on Natural Language Processing: Long Papers. Taipei, Taiwan, November 27 – December 1, 2017*, pp. 273–283.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman & Eytan Ruppin (2002): Placing search in context: The concept revisited. In: *ACM Transactions on Information Systems*, 20(1): 116–131.

John Rupert Firth (1968): A synopsis of Linguistic Theory, 1930–1955. In: *Studies in linguistic analysis*, pp. 1–32. Basil Blackwell. [No editor given. Reprint, first published 1957].

Elke Frietsch & Christina Herkommer (eds.) (2015): *Nationalsozialismus und Geschlecht. Zur Politisierung und Ästhetisierung von Körper, »Rasse« und Sexualität im »Dritten Reich« und nach 1945*. transcript-Verlag.

Aileen Fyfe, Julie McDougall-Waters & Noah Moxham (2015): 350 years of scientific periodicals. In: *Notes and Records of the Royal Society*, 69(3): 227–239.

Kata Gábor, Haïfa Zargayouna, Isabelle Tellier, Davide Buscaldi & Thierry Charnois (2017): Exploring Vector Spaces for Semantic Relations. In: *EMNLP 2017 — Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, September 7–11, 2017*, pp. 1814–1823.

Pablo Gamallo (2017): Comparing explicit and predictive distributional semantic models endowed with syntactic contexts. In: *Language Resources and Evaluation*, 51(3): 727–743.

Pablo Gamallo (2018): Evaluation of Distributional Models with the Outlier Detection Task. In: *SLATE 2018 — 7th Symposium on Languages, Applications and Technologies. Guimarães, Portugal, June 21–22, 2018*, pp. 13:1–13:8. `http://drops.dagstuhl.de/opus/volltexte/2018/9271` [Accessed May 28th 2019].

Pablo Gamallo & Stefan Bordag (2011): Is singular value decomposition useful for word similarity extraction? In: *Language Resources and Evaluation*, 45(2): 95–119.

Pablo Gamallo, Iván Rodríguez-Torres & Marcos Garcia (2018): Distributional semantics for diachronic search. In: *Computers & Electrical Engineering*, 65: 438–448.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky & James Zou (2018): Word embeddings quantify 100 years of gender and ethnic stereotypes. In: *Proceedings of the National Academy of Sciences*, 115(16): E3635–E3644.

Dirk Geeraerts (2010): *Theories of Lexical Semantics*. Oxford University Press.

Michel Généreux, Bryor Snejfella & Marta Maslej (2017): Big data in psychology: Using word embeddings to study theory-of-mind. In: *IEEE BigData 2017 — Proceedings of the 2017 IEEE International Conference on Big Data. Boston, MA, USA, December 11–14, 2017*, pp. 4747–4749.

Alexander Geyken (2013): Wege zu einem historischen Referenzkorpus des Deutschen: das Projekt Deutsches Textarchiv. In: *Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens „Altägyptisches Wörterbuch“ an der Berlin-Brandenburgischen Akademie der Wissenschaften. December 12–13, 2011*, pp. 221–234. URN: urn:nbn:de:kobv:b4-opus-24424.

Alexander Geyken & Thomas Gloning (2015): A living text archive of 15th-19th-century German. Corpus strategies, technology, organization. In: Jost Gippert & Ralf Gehrke (eds.) *Historical Corpora. Challenges and Perspectives*, pp. 165–180. Narr.

John R. Gilbert, Cleve Moler & Robert Schreiber (1992): Sparse Matrices in MATLAB: Design and Implementation. In: *SIAM Journal on Matrix Analysis and Applications*, 13(1): 333–356.

Alex Gittens, Dimitris Achlioptas & Michael W. Mahoney (2017): Skip-Gram – Zipf + Uniform = Vector Additivity. In: *ACL 2017 — Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Long Papers. Vancouver, Canada, July 30 – August 4, 2017*, pp. 69–76.

Vincent E. Giuliano (1963): Analog Networks for Word Association. In: *IEEE Transactions on Military Electronics*, MIL-7(2 & 3): 221–234.

Vincent E. Giuliano (1965): The Interpretation of Word Associations. In: Mary Elizabeth Stevens, Vincent E. Giuliano & Laurence B. Heilprin (eds.) *Statistical Association Methods For Mechanized Documentation. Symposium Proceedings. Washington, DC, USA, March 17, 1964*, pp. 25–32. `http://nvlpubs.nist.gov/nistpubs/Legacy/MP/nbsmiscellaneouspub269.pdf` [Accessed May 28th 2019].

Vincent E. Giuliano & P. E. Jones (1962): *Linear associative information retrieval.* Technical report, later appeared in “Vistas in Information Handling”. `http://www.dtic.mil/dtic/tr/fulltext/u2/290313.pdf` [Accessed May 28th 2019].

Ian Goodfellow, Yoshua Bengio & Aaron Courville (2016): *Deep*

Learning. MIT Press.

Gregory Grefenstette (1994): *Exploration in Automatic Thesaurus Discovery*. Kluwer.

Jacob Grimm & Wilhelm Grimm (eds.) (1999a): *Deutsches Wörterbuch von Jacob und WiIlhelm Grimm, Band 10, H – Juzen*. Deutscher Taschenbuch Verlag. [Facsimile of 1877 1st edition].

Jacob Grimm & Wilhelm Grimm (eds.) (1999b): *Deutsches Wörterbuch von Jacob und WiIlhelm Grimm, Band 14, R – Schiefe*. Deutscher Taschenbuch Verlag. [Facsimile of 1893 1st edition].

Kristina Gulordava & Marco Baroni (2011): A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In: *GEMS 2011 — Proceedings of the Workshop on GEometrical Models of Natural Language Semantics @ EMNLP 2011. Edinburgh, UK, July 31, 2011*, pp. 67–71.

Iryna Gurevych (2005): Using the Structure of a Conceptual Network in Computing Semantic Relatedness. In: *IJCNLP 2005 — Proceedings of the Second International Joint Conference on Natural Language Processing. Jeju Island, Korea, October 11–13, 2005*, pp. 767–778.

Susanne Haaf (2016): Corpus Analysis based on Structural Phenomena in Texts: Exploiting TEI Encoding for Linguistic Research. In: *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia, May 23–28, 2016*, pp. 4365–4372.

Susanne Haaf, Alexander Geyken & Frank Wiegand (2015): The DTA "Base Format": A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources. In: *Journal of the Text Encoding Initiative*, 8 (December 2014 – December 2015 : Selected Papers from the 2013 TEI Conference). `http://jtei.revues.org/1114` [Accessed May 28th 2019].

Susanne Haaf & Christian Thomas (2016): Die Historischen Korpora des Deutschen Textarchivs als Grundlage für sprachgeschichtliche Forschungen. In: Volker Harm, Holger Runow & Levke Schiwek (eds.) *Sprachgeschichte des Deutschen: Positionierungen in Forschung, Studium, Schule*, pp. 217–234. Hirzel.

Heidi Hahn, Beate Laudenberg & Heidi Rösch (eds.) (2015): *»Wörter raus!?« Zur Debatte um eine diskriminierungsfreie Sprache im Kinderbuch*. Beltz.

Udo Hahn, Franz Matthies, Erik Faessler & Johannes Hellrich (2016):

UIMA-Based JCoRe 2.0 Goes GitHub and Maven Central — State-of-the-Art Software Resource Engineering and Distribution of NLP Pipelines. In: *LREC 2016 — Proceedings of the Tenth International Conference on Language Resources and Evaluation. Portorož, Slovenia, May 23–28, 2016*, pp. 2502–2509.

Nathan Halko, Per-Gunnar Martinsson & Joel A. Tropp (2011): Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. In: *SIAM Review*, 53(2): 217–288.

William L. Hamilton, Kevin Clark, Jure Leskovec & Dan Jurafsky (2016a): Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora. In: *EMNLP 2016 — Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, TX, USA, November 1–5, 2016*, pp. 595–605.

William L. Hamilton, Jure Leskovec & Dan Jurafsky (2016b): Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. In: *EMNLP 2016 — Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, TX, USA, November 1–5, 2016*, pp. 2116–2121.

William L. Hamilton, Jure Leskovec & Dan Jurafsky (2016c): Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In: *ACL 2016 — Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Long Papers. Berlin, Germany, August 7–12, 2016*, pp. 1489–1501.

Zellig S. Harris (1954): Distributional Structure. In: *WORD*, 10(2–3): 146–162.

F. Heimerl & M. Gleicher (2018): Interactive Analysis of Word Vector Embeddings. In: *Computer Graphics Forum*, 37(3): 253–265.

Johannes Hellrich, Sven Buechel & Udo Hahn (2018a): JeSemE: a Website for Exploring Diachronic Changes in Word Meaning and Emotion. In: *COLING 2018 — Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations. Santa Fe, NM, USA, August 20–26, 2018*, pp. 10–14.

Johannes Hellrich, Sven Buechel & Udo Hahn (2019a): Modeling Word Emotion in Historical Language: Quantity Beats Supposed Stability in Seed Word Selection. In: *LaTeCH-CLfL 2019 — Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and*

Literature @ NAACL 2019. Minneapolis, MN, USA, June 7, 2019, pp. 1–11.

Johannes Hellrich & Udo Hahn (2016a): An Assessment of Experimental Protocols for Tracing Changes in Word Semantics Relative to Accuracy and Reliability. In: *LaTeCH 2016 — Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities @ ACL2016. Berlin, Germany, August 11, 2016*, pp. 111–117.

Johannes Hellrich & Udo Hahn (2016b): Bad company—Neighborhoods in neural embedding spaces considered harmful. In: *COLING 2016 — Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan, December 11–16, 2016*, pp. 2785–2796.

Johannes Hellrich & Udo Hahn (2016c): Measuring the dynamics of lexico-semantic change since the German Romantic period. In: *Digital Humanities 2016 — Conference Abstracts of the 2016 Conference of the Alliance of Digital Humanities Organizations (ADHO). Kraków, Poland, 11–16 July 2016*, pp. 545–547.

Johannes Hellrich & Udo Hahn (2017a): Don't Get Fooled by Word Embeddings: better Watch their Neighborhood. In: *Digital Humanities 2017 — Conference Abstracts of the 2017 Conference of the Alliance of Digital Humanities Organizations (ADHO). Montréal, Quebec, Canada, August 8–11, 2017*, pp. 250–252.

Johannes Hellrich & Udo Hahn (2017b): Exploring Diachronic Lexical Semantics with JeSemE. In: *ACL 2017 — Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Vancouver, Canada, July 30 – August 4, 2017*, pp. 31–36.

Johannes Hellrich, Bernd Kampe & Udo Hahn (2019b): The Influence of Down-Sampling Strategies on SVD Word Embedding Stability. In: *RepEval 2019 — Proceedings of the Third Workshop on Evaluating Vector Space Representations for NLP @ NAACL 2019. Minneapolis, MN, USA, June 6, 2019*, pp. 18–26.

Johannes Hellrich, Franz Matthies & Udo Hahn (2017): UIMA als Plattform für die nachhaltige Software-Entwicklung in den Digital Humanities. In: *DHd 2017 — Digitale Nachhaltigkeit, Konferenzabstracts. Bern, Switzerland, February 13–18, 2017*, pp. 279–281.

Johannes Hellrich & Christoph Rzymski (2019): Computational De-

tection of Medieval References in Metal. In: Ruth Barratt-Peacock & Ross Hagen (eds.) *Medievalism and Metal Music Studies: Throwing Down the Gauntlet*. Emerald. [To appear].

Johannes Hellrich, Alexander Stöger & Udo Hahn (2018b): Wenn der Funke überspringt — Word Embeddings im Dienst der Wissenschaftsgeschichte. In: *DHd 2018 — Kritik der digitalen Vernunft, Konferenzabstracts. Köln, Germany, February 26 – March 2, 2018*, pp. 331–335.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup & David Meger (2018): Deep Reinforcement Learning That Matters. In: *AAAI 2018 — The Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans, LA, USA, February 2–7, 2018*, pp. 3207–3214.

Felix Hill, Roi Reichart & Anna Korhonen (2014): SimLex-999: Evaluating semantic models with (Genuine) similarity estimation. In: *Computational Linguistics*, 41(4): 665–695.

Martin Hilpert (2007): Distinctive collexeme analysis and diachrony. In: *Corpus Linguistics and Linguistic Theory*, 2(2): 243–256.

Martin Hilpert & Stefan Th. Gries (2009): Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. In: *Literary and Linguistic Computing*, 24(4): 385–401.

Martin Hilpert & Florent Perek (2015): Meaning change in a petri dish: constructions, semantic vector spaces, and motion charts. In: *Linguistics Vanguard*, 1(1): 339–350.

Geoffrey E. Hinton (1986): Learning distributed representations of concepts. In: *CogSci 1986 — Proceedings of the Eighth Annual Conference of the Cognitive Science Society. Amherst, MA, USA, August 15–17, 1986*, pp. 1–12.

Oliver Hochadel (2006): The Business of Experimental Physics: Instrument Makers and Itinerant Lecturers in the German Enlightenment. In: *Science & Education*, 16(6): 525–537.

Hans Henrich Hock (1991): *Principles of historical linguistics*. Mouton de Gruyter, 2nd edition.

Heide Hollmer (2010): Schlegel, Johann Elias. In: Bernd Lutz & Benedikt Jeßing (eds.) *Metzler Autorenlexikon. Deutschsprachige Dichter und Schriftsteller vom Mittelalter bis zur Gegenwart*, p. 683. J.B. Metzler, 4th edition.

R. W. Home (2008): Mechanics and Experimental Physics. In:

Roy Porter (ed.) *The Cambridge History of Science, Volume 4: Eighteenth-Century Science*, pp. 354–374. Cambridge University Press.

Dirk Hovy & Christoph Purschke (2018): Capturing Regional Variation with Distributed Place Representations and Geographic Retrofitting. In: *EMNLP 2018 — Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, October 31 – November 4, 2018*, pp. 4383–4394.

Susan Hunston (2002): *Corpora in Applied Linguistics*. Cambridge University Press.

Matthew Hutson (2018): Artificial intelligence faces reproducibility crisis. In: *Science*, 359(6377): 725–726.

Peter Ivie & Douglas Thain (2018): Reproducibility in Scientific Computing. In: *ACM Computing Surveys*, 51(3): 63:1–63:36.

Paul Jaccard (1912): The distribution of the flora in the alpine zone. In: *New Phytologist*, XI(2): 37–50. [Translation of 1901 article].

Ilse Jahn (1994): On the origin of romantic biology and its further development at the university of Jena between 1790 and 1850. In: Stefano Poggi & Maurizio Bossi (eds.) *Romanticism in Science. Science in Europe, 1790–1840*, pp. 75–89. Kluwer Acadmic Publishers.

Kokil Jaidka, Niyati Chhaya & Lyle H. Ungar (2018): Diachronic degradation of language models: Insights from social media. In: *ACL 2018 — Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Short Papers. Melbourne, Australia, July 15–20, 2018*, pp. 195–200.

Adam Jatowt & Kevin Duh (2014): A framework for analyzing semantic change of words across time. In: *JCDL 2014 — Proceedings of the 14th ACM-IEEE-CS Joint Conference on Digital Libraries. London, U.K., September 8–12, 2014*, pp. 229–238.

Eun Seo Jo (2016): Diplomatic history by data. Understanding Cold War foreign policy ideology using networks and NLP. In: *Digital Humanities 2016 — Conference Abstracts of the 2016 Conference of the Alliance of Digital Humanities Organizations (ADHO). 'Digital Identities: The Past and the Future'. Kraków, Poland, 11–16 July 2016*, pp. 582–585.

Matthew L. Jockers (2013): *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.

Edgar Jones (2010): Google books as a general research collection.

In: *Library Resources & Technical Services*, 54(2): 77–89.

Daniel Jurafsky & James H. Martin (2009): *Speech and Language Processing*. Prentice Hall, 2nd edition.

David Jurgens & Keith Stevens (2009): Event Detection in Blogs using Temporal Random Indexing. In: *eETTs 2009 — Proceedings of the Workshop on Events in Emerging Text Types @ RANLP 2009. Borovets, Bulgaria, September 17, 2009*, pp. 9–16.

Bryan Jurish (2013): Canonicalizing the Deutsches Textarchiv. In: Ingelore Hafemann (ed.) *Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens „Altägyptisches Wörterbuch" an der Berlin-Brandenburgischen Akademie der Wissenschaften. December 12–13, 2011*, pp. 235–244. BBAW.

Bryan Jurish (2015): DiaCollo: On the trail of diachronic collocations. In: *Proceedings of the CLARIN Annual Conference 2015. Book of Abstracts. Wrocław,, Poland, 14-16 October, 2015*, pp. 28–31.

Pentti Kanerva, Jan Kristoferson & Anders Holst (2000): Random indexing of text samples for latent semantic analysis. In: *CogSci 2000 — Proceedings of the 22nd Annual Conference of the Cognitive Science Society. Philadelphia, PA, USA, August 13–15, 2000*, p. 1036.

Tom Kenter, Melvin Wevers, Pim Huijnen & Maarten de Rijke (2015): Ad hoc monitoring of vocabulary shifts over time. In: *CIKM 2015 — Proceedings of the 24th ACM International Conference on Information and Knowledge Management. Melbourne, Australia, October 19–23, 2015*, pp. 1191–1200.

Hannah Kermes, Stefania Degaetano-Ortlieb, Ashraf Khamis, Jörg Knappen & Elke Teich (2016): The Royal Society Corpus: From uncharted data to corpus. In: *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia, May 23–28, 2016*, pp. 1928–1931.

Evgeny Kim, Sebastian Padó & Roman Klinger (2017): Investigating the Relationship between Literary Genres and Emotional Plot Development. In: *LaTeCH-CLfL 2017 — Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature @ ACL 2017. Vancouver, Canada, August 4, 2017*, pp. 17–26.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde & Slav Petrov

(2014): Temporal analysis of language through neural language models. In: *LACS 2014 — Proceedings of the Workshop on Language Technologies and Computational Social Science @ ACL 2014. Baltimore, MD, USA, June 26, 2014*, pp. 61–65.

Manfred Klenner & Udo Hahn (1994): Concept Versioning: A Methodology for Tracking Evolutionary Concept Drift in Dynamic Concept Systems. In: *ECAI 1994 — Proceedings of the 11th European Conference on Artificial Intelligence. Amsterdam, The Netherlands, August 8-12, 1994*, pp. 473–477.

David Knight (1990): Romanticism and the sciences. In: Andrew Cunningham & Nichola Jardine (eds.) *Romanticism and the sciences*, pp. 13–24. Cambridge University Press.

Thomas Kober, Julie Weeds, John Wilkie, Jeremy Reffin & David Weir (2017): One Representation per Word - Does it make Sense for Composition? In: *SENSE 2017 — Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications @ EACL 2017. Valencia, Spain, April 4, 2017*, pp. 79–90.

Thomas Kohnen (2006): Historical Corpus Linguistics: Perspectives on English diachronic corpora. In: *Anglistik*, 17(2): 73–91.

Matthew B. Koll (1979): WEIRD: An Approach to Concept-based Information Retrieval. In: *SIGIR Forum*, 13(4): 32–50.

Michal Konkol, Tomáš Brychcín, Michal Nykl & Tomáš Hercig (2017): Geographical Evaluation of Word Embeddings. In: *IJCNLP 2017 — Proceedings of the Eighth International Joint Conference on Natural Language Processing: Long Papers. Taipei, Taiwan, November 27 – December 1, 2017*, pp. 224–232.

Maximilian Köper & Sabine Schulte im Walde (2016): Automatically generated affective norms of abstractness, arousal, imageability and valence for 350,000 German lemmas. In: *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia, May 23–28, 2016*, pp. 2595–2598.

Alexander Koplenig (2017): The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets—Reconstructing the composition of the German corpus in times of WWII. In: *Digital Scholarship in the Humanities*, 32(1): 169–188.

Detlef Kremer & Andreas B. Kilcher (2015): *Romantik. Lehrbuch*

Germanistik. J.B. Metzler, 4th edition.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi & Steven Skiena (2015): Statistically significant detection of linguistic change. In: *WWW 2015 — Proceedings of the 24th International Conference on World Wide Web: Technical Papers. Florence, Italy, May 18–22, 2015*, pp. 625–635.

Vivek Kulkarni, Bryan Perozzi & Steven Skiena (2016): Freshman or fresher? Quantifying the geographic variation of language in online social media. In: *ICWSM-16 — Proceedings of the 10th International AAAI Conference on Web and Social Media. Cologne, Germany, May 17–20, 2016*, pp. 615–618.

Claudia Kunze & Lothar Lemnitzer (2002): GermaNet: Representation, visualization, application. In: *LREC 2002 — Proceedings of the 3rd International Conference on Language Resources and Evaluation. Las Palmas, Canary Islands, Spain, 27 May – June 2, 2002*, pp. 1485–1491.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski & Erik Velldal (2018): Diachronic word embeddings and semantic shifts: a survey. In: *COLING 2018 — Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, NM, USA, August 20–26, 2018*, pp. 1384–1397.

Andrey Kutuzov, Erik Velldal & Lilja Øvrelid (2017): Tracing armed conflicts with diachronic word embedding models. In: *EventStory 2017 — Proceedings of the Events and Stories in the News Workshop @ ACL 2017. Vancouver, Canada, August 4, 2017*, pp. 31–36.

Jouni-Matti Kuukkanen (2008): Making sense of conceptual change. In: *History and Theory*, 47(3): 351–372.

Tim F. LaHaye & Jerry B. Jenkins (2007): *Kingdom come : the final victory*. Thorndike Press.

Thomas K. Landauer & Susan T. Dumais (1997): A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. In: *Psychological Review*, 104(2): 211–240.

Rémi Lebret & Ronan Collobert (2015): Rehabilitation of Count-Based Models for Word Vector Representations. In: *CICLing 2015 — Computational Linguistics and Intelligent Text Processing, 16th International Conference, Part I. Cairo, Egypt, April 14–20, 2015*, pp. 417–429.

Yann LeCun, Yoshua Bengio & Geoffrey E. Hinton (2015): Deep learning. In: *Nature*, 521(7553): 436–444.

Guang-He Lee & Yun-Nung Chen (2017): MUSE: Modularizing Unsupervised Sense Embeddings. In: *EMNLP 2017 — Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, September 7–11, 2017*, pp. 327–337.

Jakob Michael Reinhold Lenz (1774): *Anmerkungen übers Theater, nebst angehängten übersetzten Stück Shakespears.* Weigand, 1st edition.

Omer Levy & Yoav Goldberg (2014a): Dependency-based word embeddings. In: *ACL 2014 — Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Short Papers. Baltimore, MD, USA, June 22–27, 2014*, pp. 302–308.

Omer Levy & Yoav Goldberg (2014b): Linguistic Regularities in Sparse and Explicit Word Representations. In: *CoNLL 2014 — Proceedings of the Eighteenth Conference on Computational Natural Language Learning. Baltimore, MD, USA, June 26–27, 2014*, pp. 171–180.

Omer Levy & Yoav Goldberg (2014c): Neural Word Embedding as Implicit Matrix Factorization. In: *NIPS 2014 — Advances in Neural Information Processing Systems 27. Montréal, Quebec, Canada, December 8–13, 2014*, pp. 2177–2185.

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

P. A. W. Lewis, P. B. Baxendale & J. L. Bennett (1967): Statistical Discrimination of the Synonymy/Antonymy Relationship Between Words. In: *Journal of the ACM*, 14(1): 20–44.

Ying Li, Tomas Engelthaler, Cynthia S. Q. Siew & Thomas T. Hills (2019): The Macroscope: A tool for examining the historical structure of language. In: *Behavior Research Methods.* `https://doi.org/10.3758/s13428-018-1177-6` [Accessed May 28th 2019].

Georg Christoph Lichtenberg & Johann Christian Polykarp Erxleben (1787): *Anfangsgründe der Naturlehre.* Dietrich, 4th edition.

Dekang Lin (1998): Automatic retrieval and clustering of similar words. In: *COLING-ACL 1998 — Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics & 17th*

International Conference on Computational Linguistics. Montréal, Quebec, Canada, August 10–14, 1998, volume 2, pp. 768–774.

Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman & Slav Petrov (2012): Syntactic annotations for the Google Books Ngram Corpus. In: *ACL 2012 — Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Jeju Island, Korea, July 10, 2012*, pp. 169–174.

Charles H. Lippy (2010): Politics. In: Amanda Porterfield & John Corrigan (eds.) *Religion in American history*, pp. 249–265. Wiley-Blackwell.

Kevin Lund & Curt Burgess (1996): Producing high-dimensional semantic spaces from lexical co-occurrence. In: *Behavior Research Methods, Instruments, & Computers*, 28(2): 203–208.

Kevin Lund, Curt Burgess & Ruth Ann Atchley (1995): Semantic and associative priming in high-dimensional semantic space. In: *CogSci 1995 — Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society. Pittsburgh, Pennsylvania, July 22–25, 1995*, pp. 660–665.

John Lyons (1996): *Linguistic Semantics: An Introduction*. Cambridge University Press. Reprint of 1995 edition.

Chris Manning & Hinrich Schütze (1999): *Foundations of Statistical Natural Language Processing*. MIT Press.

Christopher D. Manning (2015): Computational Linguistics and Deep Learning. In: *Computational Linguistics*, 41(4): 701–707.

M. E. Maron & J. L. Kuhns (1960): On Relevance, Probabilistic Indexing and Information Retrieval. In: *Journal of the ACM*, 7(3): 216–244.

Tony McEnery & Andrew Wilson (1996): *Corpus Linguistics*. Edinburgh University Press.

Jill P. Mesirov (2010): Accessible Reproducible Research. In: *Science*, 327(5964): 415–416.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak & Erez Lieberman Aiden (2011): Quantitative analysis of culture using millions of digitized books. In: *Science*, 331(6014): 176–182.

Rada Mihalcea & Vivi Nastase (2012): Word epoch disambiguation:

Finding how words change over time. In: *ACL 2012 — Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers. Jeju Island, Korea, July 8–14, 2012*, pp. 259–263.

Tomas Mikolov, Kai Chen, Greg Corrado & Jeffrey Dean (2013a): Efficient estimation of word representations in vector space. In: *ICLR 2013 — Workshop Proceedings of the International Conference on Learning Representations. Scottsdale, AZ, USA, May 2–4, 2013*. `http://arxiv.org/abs/1301.3781` [Accessed May 28th 2019].

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch & Armand Joulin (2018): Advances in Pre-Training Distributed Word Representations. In: *LREC 2018 — Proceedings of the Eleventh International Conference on Language Resources and Evaluation. Miyazaki, Japan, May 7–12, 2018*, pp. 52–55.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado & Jeffrey Dean (2013b): Distributed representations of words and phrases and their compositionality. In: *NIPS 2013 — Advances in Neural Information Processing Systems 26. Lake Tahoe, NV, USA, December 5-10, 2013*, pp. 3111–3119.

Tomas Mikolov, Wen-tau Yih & Geoffrey Zweig (2013c): Linguistic regularities in continuous space word representations. In: *NAACL-HLT 2013 — Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, GA, USA, 9–14 June 2013*, pp. 746–751.

George A. Miller (1995): WordNet: A Lexical Database for English. In: *Communications of the ACM*, 38(11): 39–41.

A. D. Mills (ed.) (2011): *A Dictionary of British Place Names*. Oxford University Press. 2012 Online Version, DOI: 10.1093/acref/9780199609086.001.0001.

David Mimno (2012): Computational historiography: Data mining in a century of classics journals. In: *Journal on Computing and Cultural Heritage*, 5(1). Article 3.

David Mimno & Laure Thompson (2017): The strange geometry of skip-gram with negative sampling. In: *EMNLP 2017 — Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, September 7–11, 2017*, pp. 2873–2878.

Sunny Mitra, Ritwik Mitra, Suman Kalyan Maity, Martin Riedl,

Chris Biemann, Pawan Goyal & Animesh Mukherjee (2015): An automatic approach to identify word sense changes in text media across timescales. In: *Natural Language Engineering*, 21(5): 773–798.

Sunny Mitra, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee & Pawan Goyal (2014): That's sick dude!: Automatic identification of word sense change across different timescales. In: *ACL 2014 — Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Long Papers. Baltimore, MD, USA, June 22–27, 2014*, pp. 1020–1029.

Saif Mohammad (2008): *Measuring Semantic Distance using Distributional Profiles of Concepts*. Ph.D. thesis, University of Toronto. `http://hdl.handle.net/1807/11238` [Accessed May 28th 2019].

Franco Moretti (2013): *Distant Reading*. Verso.

Iwan Rhys Morus (1998): *Frankenstein's Children. Electricity, Exhibition, And Experiment in Early-Nineteenth-Century London*. Princeton University Press.

Iwan Rhys Morus (2011): *Shocking Bodies: Life, Death & Electricity in Victorian England*. The History Press.

Avo Muromägi, Kairit Sirts & Sven Laur (2017): Linear Ensembles of Word Embedding Models. In: *NoDaLiDa 2017 — Proceedings of the 21st Nordic Conference on Computational Linguistics. Gothenburg, Sweden, May 22–24, 2017*, pp. 96–104.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos & Andrew McCallum (2014): Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space. In: *EMNLP 2014 — Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, October 25–29, 2014*, pp. 1059–1069.

John von Neumann (1963): Various techniques used in connection with random digits, Summary written by George E. Forsythe. In: *John von Neumann, Collected Works*, volume 5, pp. 768–770. Pergamon Press. [Reprint of 1951 article].

Susanne Niemeier (2003): Straight from the heart: metonymic and metaphorical explorations. In: Antonio Barcelona (ed.) *Metaphor and Metonymy at the Crossroads: A Cognitive Perspective*, pp. 195–213. Mouton de Gruyter.

Yoshiki Niwa & Yoshihiko Nitta (1994): Co-occurrence Vectors from Corpora vs. Distance Vectors from Dictionaries. In: *COLING*

1994 — Proceedings of the 15th Conference on Computational Linguistics: Volume 1. Kyoto, Japan, August 5–9, 1994, pp. 304–309.

Frederick O'Brien (1921): *Mystic Isles of the South Seas.*

Open Science Collaboration (2015): Estimating the reproducibility of psychological science. In: *Science*, 349(6251): aac4716.

Charles E. Osgood (1952): The nature and measurement of meaning. In: *Psychological Bulletin*, 49(3): 197–237.

Charles E. Osgood (1953): *Method and theory in experimental psychology.* Oxford University Press.

Charles E. Osgood, George J. Suci & Percy H. Tannenbaum (1957): *The Measurement of Meaning.* University of Illinois Press.

Arvid Österlund, David Ödling & Magnus Sahlgren (2015): Factorization of Latent Variables in Distributional Semantic Models. In: *EMNLP 2015 — Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 17–21 September 2015*, pp. 227–231.

O. P. O'Sullivan, R. M. Duffy & B. D. Kelly (2017): Culturomics and the history of psychiatry: testing the Google Ngram method. In: *Irish Journal of Psychological Medicine.* `https://doi.org/10.1017/ipm.2017.37` [Accessed May 28th 2019].

Muntsa Padró, Marco Idiart, Aline Villavicencio & Carlos Ramisch (2014): Comparing Similarity Measures for Distributional Thesauri. In: *LREC 2014 — Proceedings of the 9th International Conference on Language Resources and Evaluation. Reykjavik, Iceland, May 26-31, 2014*, pp. 2694–2971.

Elisavet Palogiannidi, Polychronis Koutsakis, Elias Iosif & Alexandros Potamianos (2016): Affective Lexicon Creation for the Greek Language. In: *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia, May 23–28, 2016*, pp. 2867–2872.

Bo Pang & Lillian Lee (2008): Opinion mining and sentiment analysis. In: *Foundations and Trends in Information Retrieval*, 2(1-2): 1–135.

Joohee Park & Sung-hyon Myaeng (2017): A Computational Study on Word Meanings and Their Distributed Representations via Polymodal Embedding. In: *IJCNLP 2017 — Proceedings of the Eighth International Joint Conference on Natural Language Processing: Long Papers. Taipei, Taiwan, November 27 – December*

1, 2017, pp. 214–223.

Kevin Patel & Pushpak Bhattacharyya (2017): Towards Lower Bounds on Number of Dimensions for Word Embeddings. In: *IJCNLP 2017 — Proceedings of the Eighth International Joint Conference on Natural Language Processing: Long Papers. Taipei, Taiwan, November 27 – December 1, 2017*, pp. 31–36.

Eitan Adam Pechenick, Christopher M. Danforth & Peter Sheridan Dodds (2015): Characterizing the Google Books Corpus: strong limits to inferences of socio-cultural and linguistic evolution. In: *PLoS One*, 10(10): e0137041.

Ted Pedersen, Siddharth Patwardhan & Jason Michelizzi (2004): WordNet::Similarity - Measuring the Relatedness of Concepts. In: *HLT-NAACL 2004 — Proceedings of the 2004 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: System Demonstrations. Boston, MA, USA, May 2–7, 2004*, pp. 38–41.

Jeffrey Pennington, Richard Socher & Christopher D. Manning (2014): GloVe: Global vectors for word representation. In: *EMNLP 2014 — Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, October 25–29, 2014*, pp. 1532–1543.

Florent Perek (2014): Vector spaces for historical linguistics: Using distributional semantics to study syntactic productivity in diachrony. In: *ACL 2014 — Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Short Papers. Baltimore, MD, USA, June 22–27, 2014*, pp. 309–314.

Seamus Perry (1998): Romanticism: The Brief History of a Concept. In: Duncan Wu (ed.) *A Companion to Romanticism*, pp. 3–11. Blackwell.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher T. Clark, Kenton Lee & Luke S. Zettlemoyer (2018): Deep contextualized word representations. In: *NAACL-HLT 2018 — Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Long Papers. New Orleans, Louisiana, USA, June 1-6, 2018*, pp. 2227–2237.

Bénédicte Pierrejean & Ludovic Tanguy (2018a): Étude de la reproductibilité des word embeddings: repérage des zones stables et instables dans le lexique. In: Pascale Sébillot & Vincent

Claveau (eds.) *TALN 2018 — Actes de la 25ème conférence sur le Traitement Automatique des Langues Naturelles. Rennes, France, 14-18 Mai, 2018.*, volume 1: Articles longs, articles courts de TALN, pp. 33–46.

Bénédicte Pierrejean & Ludovic Tanguy (2018b): Towards Qualitative Word Embeddings Evaluation: Measuring Neighbors Variation. In: *NAACL-HLT 2018 — Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop. New Orleans, LA, USA, June 2–4, 2018*, pp. 32–39.

Steven Pinker (1994): The Game of the Name. In: *The New York Times*. April 5th. `https://stevenpinker.com/files/pinker/files/1994_04_03_newyorktimes.pdf` [Accessed May 28th 2019].

Christian Pölitz, Thomas Bartz, Katharina Morik & Angelika Störrer (2015): Investigation of Word Senses over Time Using Linguistic Corpora. In: *TSD 2015 — Text, Speech, and Dialogue. Proceedings of the 18th International Conference. Pilsen, Czech Republic, September 14–17, 2015*, pp. 191–198.

Octavian Popescu & Carlo Strapparava (2013): Behind the Times: Detecting Epoch Changes using Large Corpora. In: *IJCNLP 2013 — Proceedings of the Sixth International Joint Conference on Natural Language Processing. Nagoya, Japan, October 14–19, 2013*, pp. 347–355.

Octavian Popescu & Carlo Strapparava (2015): SemEval 2015, Task 7: Diachronic Text Evaluation. In: *SemEval 2015 — Proceedings of the 9th International Workshop on Semantic Evaluation. Denver, CO, USA, June 4–5, 2015*, pp. 870–878.

Daniel Preoţiuc-Pietro, P. K. Srijith, Mark Hepple & Trevor Cohn (2016): Studying the Temporal Dynamics of Word Co-occurrences: An Application to Event Detection. In: *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia, May 23–28, 2016*, pp. 4380–4387.

Roy Rada, Hafedh Mili, Ellen Bicknell & Maria Blettner (1989): Development and application of a metric on semantic nets. In: *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1): 17–30.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich & Shaul Markovitch (2011): A word at a time: Computing word relatedness using temporal semantic analysis. In: *WWW 2011 — Proceedings*

of the 20th international conference on World Wide Web. Hyderabad, India, March 28 – April 1, 2011, pp. 337–346.

Gabriel Recchia, Ewan Jones, Paul Nulty, John Regan & Peter de Bolla (2017): Tracing Shifting Conceptual Vocabularies Through Time. In: *Knowledge Engineering and Knowledge Management. EKAW 2016 Satellite Events, EKM and Drift-an-LOD. Bologna, Italy, November 19–23, 2016, Revised Selected Papers*, pp. 19–28. Springer.

Nils Reimers & Iryna Gurevych (2017): Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In: *EMNLP 2017 — Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, September 7–11, 2017*, pp. 338–348.

Joseph Reisinger & Raymond J. Mooney (2010): Multi-Prototype Vector-Space Models of Word Meaning. In: *NAACL-HLT 2010 — Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Los Angeles, California, June 1–6, 2010*, pp. 109–117.

Martin Riedl & Chris Biemann (2013): Scaling to Large^3 Data: An Efficient and Effective Method to Compute Distributional Thesauri. In: *EMNLP 2013 — Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, WA, USA, October 18–21, 2013*, pp. 884–890.

Martin Riedl, Richard Steuer & Chris Biemann (2014): Distributed Distributional Similarities of Google Books over the Centuries. In: *LREC 2014 — Proceedings of the 9th International Conference on Language Resources and Evaluation. Reykjavik, Iceland, May 26-31, 2014*, pp. 1401–1405.

Martina Astrid Rodda, Marco S. G. Senaldi & Alessandro Lenci (2016): Panta Rei: Tracking Semantic Change with Distributional Semantics in Ancient Greek. In: *CLiC-it 2016 — Proceedings of Third Italian Conference on Computational Linguistics. Napoli, Italy, December 5-7, 2016.* `http://ceur-ws.org/Vol-1749/paper46.pdf` [Accessed May 28$^{\text{th}}$ 2019].

Alex Rosenfeld & Katrin Erk (2018): Deep Neural Models of Semantic Shift. In: *NAACL-HLT 2018 — Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Long*

Papers. New Orleans, LA, USA, June 2–4, 2018, pp. 474–484.

Sascha Rothe, Sebastian Ebert & Hinrich Schütze (2016): Ultradense Word Embeddings by Orthogonal Transformation. In: *NAACL-HLT 2016 — Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, June 12–17, 2016*, pp. 767–777.

Herbert Rubenstein & John B. Goodenough (1965): Contextual Correlates of Synonymy. In: *Communications of the ACM*, 8(10): 627–633.

Maja Rudolph & David Blei (2018): Dynamic Embeddings for Language Evolution. In: *WWW 20018 — Proceedings of the 2018 World Wide Web Conference. Lyon, France, April 23–27, 2018*, pp. 1003–1011.

Yousef Saad (2003): *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, Philadelphia/PA, 2nd edition.

Eyal Sagi, Stefan Kaufmann & Brady Clark (2009): Semantic Density Analysis: Comparing Word Meaning across Time and Phonetic Space. In: *GEMS 2009 — Proceedings of the GEometrical Models of Natural Language Semantics Workshop @ EACL 2009. Athens, Greece, March 31, 2009*, pp. 104–111.

Eyal Sagi, Stefan Kaufmann & Brady Clark (2012): Tracing semantic change with Latent Semantic Analysis. In: Kathryn Allan & Justyna A. Robinson (eds.) *Current Methods in Historical Semantics*, pp. 161–183. De Gruyter Mouton.

Magnus Sahlgren (2006): *The Word-Space Model — Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Stockholm University. `http://su.diva-portal.org/smash/get/diva2:189276/FULLTEXT01` [Accessed May 28th 2019].

Magnus Sahlgren & Alessandro Lenci (2016): The Effects of Data Size and Frequency Range on Distributional Semantic Models. In: *EMNLP 2016 — Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, TX, USA, November 1–5, 2016*, pp. 975–980.

G. Salton (ed.) (1971): *The SMART Rretrieval Ssystem: Eexperiments in Aautomatic Ddocument Processing*. Prentice-Hall.

G. Salton & M. E. Lesk (1971): Information analysis and dictionary

construction. In: G. Salton (ed.) *The SMART Rretrieval Ssystem: Eexperiments in Aautomatic Ddocument Processing*, chapter 6, pp. 115–142. Prentice-Hall.

G. Salton, A. Wong & C. S. Yang (1975): A Vector Space Model for Automatic Indexing. In: *Communications of the ACM*, 18(11): 613–620.

Geir Kjetil Sandve, Anton Nekrutenko, James Taylor & Eivind Hovig (2013): Ten Simple Rules for Reproducible Computational Research. In: *PLoS Computational Biology*, 9(10): e1003285.

Wayne M. Saslow (2002): *Electricity, Magnetism, and Light*. Academic Press.

Christin Schätzle, Michael Hund, Frederik L. Dennig, Miriam Butt & Daniel A. Keim (2017): HistoBankVis: Detecting language change via data visualization. In: *Proceedings of the Workshop on Processing Historical Language @ NoDaLiDa 2017. Gothenburg, Sweden, May 22, 2017*, pp. 32–39.

Hannelore Schlaffer (2010a): Goethe, Johann Wolfgang von. In: Bernd Lutz & Benedikt Jeßing (eds.) *Metzler Autorenlexikon. Deutschsprachige Dichter und Schriftsteller vom Mittelalter bis zur Gegenwart*, pp. 227–232. J.B. Metzler, 4th edition.

Heinz Schlaffer (2010b): Schlegel, August Wilhelm. In: Bernd Lutz & Benedikt Jeßing (eds.) *Metzler Autorenlexikon. Deutschsprachige Dichter und Schriftsteller vom Mittelalter bis zur Gegenwart*, pp. 680–681. J.B. Metzler, 4th edition.

Dominik Schlechtweg, Stefanie Eckmann, Enrico Santus, Sabine Schulte im Walde & Daniel Hole (2017): German in Flux: Detecting Metaphoric Change via Word Entropy. In: *CoNLL 2017 — Proceedings of the 21st Conference on Computational Natural Language Learning. Vancouver, Canada, August 3–4, 2017*, pp. 354–367.

Wilhelm Schmidt (2007): *Geschichte der deutschen Sprache. Ein Lehrbuch für das germanistische Studium*. Hirzel, 10th edition.

David S. Schmidtke, Tobias Schröder, Arthur M. Jacobs & Markus Conrad (2014): ANGST: Affective norms for German sentiment terms, derived from the affective norms for English words. In: *Behavior Research Methods*, 46(4): 1108–1118.

Tobias Schnabel, Igor Labutov, David Mimno & Thorsten Joachims (2015): Evaluation methods for unsupervised word embeddings. In: *EMNLP 2015 — Proceedings of the 2015 Conference on Empirical*

Methods in Natural Language Processing. Lisbon, Portugal, 17–21 September 2015, pp. 298–307.

Christof Schöch (2017): Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama. In: *Digital Humanities Quarterly*, 11(2). `http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html` [Accessed May 28th 2019].

Peter Schönemann (1966): A generalized solution of the orthogonal procrustes problem. In: *Psychometrika*, 31: 1–10.

Hinrich Schütze (1992a): Dimensions of Meaning. In: *Supercomputing 1992 — Proceedings of the 1992 ACM/IEEE Conference on Supercomputing. Minneapolis, MN, USA, November 16–20, 1992*, pp. 787–796.

Hinrich Schütze (1992b): Word Space. In: *NIPS 1992 — Advances in Neural Information Processing Systems 5. Denver, CO, USA, November 30 – December 3, 1992*, pp. 895–902.

Hinrich Schütze (1993): Part-of-speech Induction from Scratch. In: *ACL 1993 — Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics. Columbus, OH, USA, June 22–26 , 1993*, pp. 251–258.

Hinrich Schütze & Jan Pedersen (1993): A vector model for syntagmatic and paradigmatic relatedness. In: *Proceedings of the Ninth Annual Conference of UW Centre for the New OED and Text Research. Oxford, England, September 27–28, 1993*, pp. 104–113.

C.E. Shannon (1948): A mathematical theory of communication. In: *The Bell System Technical Journal*, 27(3): 379–423.

J. A. Simpson & E. S. C. Weiner (eds.) (1989a): *The Oxford English Dictionary. Volume XII, Hat–Intervacuum.* Clarendon Press, 2nd edition.

J. A. Simpson & E. S. C. Weiner (eds.) (1989b): *The Oxford English Dictionary. Volume XIV, Rob–Sequyle.* Clarendon Press, 2nd edition.

Frank A. Smadja & Kathleen R. McKeown (1990): Automatically Extracting and Representing Collocations for Language Generation. In: *ACL 1990 — Proceedings of the 28th Annual Meeting on Association for Computational Linguistics. Pittsburgh, PA, USA, June 6–9 , 1990*, pp. 252–259.

Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng & Christopher D. Manning (2011): Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions. In: *EMNLP*

2011 — Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, Scotland, UK, July 27–29, 2011, pp. 151–161.

Joseph Spiegel & Edward Bennett (1965): A modified statistical association procedure for automatic document content analysis and retrieval. In: Mary Elizabeth Stevens, Vincent E. Giuliano & Laurence B. Heilprin (eds.) *Statistical Association Methods For Mechanized Documentation. Symposium Proceedings. Washington, DC, USA, March 17, 1964*, pp. 47–60. `http://nvlpubs.nist.gov/nistpubs/Legacy/MP/nbsmiscellaneouspub269.pdf` [Accessed May 28th 2019].

Joan Steigerwald (2013): Rethinking Organic Vitality in Germany at the Turn of the Nineteenth Century. In: Sebastian Normandin & Charles T. Wolfe (eds.) *Vitalism and the Scientific Image in Post-Enlightenment Life Science, 1800-2010*. Springer.

David Stevens (2004): *Romanticism*. Cambridge University Press.

H. Edmund Stiles (1961): The Association Factor in Information Retrieval. In: *Journal of the ACM*, 8(2): 271–279.

Philip J. Stone & Earl B. Hunt (1963): A Computer Approach to Content Analysis: Studies Using the General Inquirer System. In: *AFIPS 1963 — Proceedings of the Spring Joint Computer Conference. Detroit, Michigan, May 21–23, 1963*, pp. 241–256.

Paul Switzer (1965): Vector images in document retrieval. In: Mary Elizabeth Stevens, Vincent E. Giuliano & Laurence B. Heilprin (eds.) *Statistical Association Methods For Mechanized Documentation. Symposium Proceedings. Washington, DC, USA, March 17, 1964*, pp. 163–171. `http://nvlpubs.nist.gov/nistpubs/Legacy/MP/nbsmiscellaneouspub269.pdf` [Accessed May 28th 2019].

Terrence Szymanski (2017): Temporal Word Analogies: Identifying Lexical Replacement with Diachronic Word Embeddings. In: *ACL 2017 — Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Short Papers. Vancouver, Canada, July 30 – August 4, 2017*, pp. 448–453.

Irma Taavitsainen (2015): Medical news in England 1665–1800 in journals for professional and lay audiences. In: Birte Böse & Lucia Kornexl (eds.) *Changing Genre Conventions in Historical English News Discourse*, pp. 135–159. John Benjamins.

Nina Tahmasebi (2018): A Study on Word2Vec on a Historical

Swedish Newspaper Corpus. In: *DHN 2018 — Proceedings of the Digital Humanities in the Nordic Countries, 3rd Conference. Helsinki, Finland, March 7–9, 2018*, pp. 25–37.

Nina Tahmasebi, Gerhard Gossen, Nattiya Kanhabua, Helge Holzmann & Thomas Risse (2012): NEER: An Unsupervised Method for Named Entity Evolution Recognition. In: *COLING 2012 — Proceedings of the 24th International Conference on Computational Linguistics: Technical Papers. Mumbai, India, December 8–15, 2012*, pp. 2553–2568.

Nina Tahmasebi & Thomas Risse (2017a): Finding Individual Word Sense Changes and their Delay in Appearance. In: *RANLP 2017 — Proceedings of the International Conference Recent Advances in Natural Language Processing. Varna, Bulgaria, September 2–8, 2017*, pp. 741–749.

Nina Tahmasebi & Thomas Risse (2017b): On the Uses of Word Sense Change for Research in the Digital Humanities. In: *TPDL 2017 — Research and Advanced Technology for Digital Libraries: 21st International Conference on Theory and Practice of Digital Libraries. Thessaloniki, Greece, September 18–21, 2017*, pp. 246–257.

Nina N. Tahmasebi (2013): *Models and Algorithms for Automatic Detection of Language Evolution. Towards Finding and Interpreting of Content in Long-Term Archives.* Ph.D. thesis, Gottfried Wilhelm Leibniz Universität Hannover. `http://edok01.tib.uni-hannover.de/edoks/e01dh13/771705034.pdf` [Accessed May 28th 2019].

Hiroya Takamura, Ryo Nagata & Yoshifumi Kawasaki (2017): Analyzing Semantic Change in Japanese Loanwords. In: *EACL 2017 — Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Long Papers. Valencia, Spain, April 3–7, 2017*, pp. 1195–1204.

Simon Tanner, Trevor Muñoz & Pich Hemy Ros (2009): Measuring Mass Text Digitization Quality and Usefulness, Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive. In: *D-Lib Magazine*, 15(7/8). `http://www.dlib.org/dlib/july09/munoz/07munoz.html` [Accessed May 28th 2019].

Paul R. Thagard (1990): Concepts and conceptual change. In: *Synthese*, 82(2): 255–274.

Roberto Theron & Laura Fontanillo (2015): Diachronic-information visualization in historical dictionaries. In: *Information Visualization*, 14(2): 111–136.

Joseph Turian, Lev Ratinov & Yoshua Bengio (2010): Word Representations: A Simple and General Method for Semi-Supervised Learning. In: *ACL 2010 — Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics: Long Papers. Uppsala, Sweden, July 11–16, 2010*, pp. 384–394.

Peter D. Turney (2001): Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In: *ECML 2001 — Proceedings of the 12th European Conference on Machine Learning. Freiburg, Germany, September 5–7, 2001*, pp. 491–502.

Peter D. Turney (2002): Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: *ACL 2002 — Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, PA, USA, July 6–12, 2002*, pp. 417–424.

Peter D. Turney & Michael L. Littman (2003): Measuring praise and criticism: Inference of semantic orientation from association. In: *ACM Transactions on Information Systems*, 21(4): 315–346.

Peter D. Turney & Patrick Pantel (2010): From frequency to meaning: Vector space models of semantics. In: *Journal of Artificial Intelligence Research*, 37(1): 141–188.

Friedrich Vollhardt (2010): Lessing, Gotthold Ephraim. In: Bernd Lutz & Benedikt Jeßing (eds.) *Metzler Autorenlexikon. Deutschsprachige Dichter und Schriftsteller vom Mittelalter bis zur Gegenwart*, pp. 497–501. J.B. Metzler, 4th edition.

Daniel D. Walker, William B. Lund & Eric K. Ringger (2010): Evaluating Models of Latent Document Semantics in the Presence of OCR Errors. In: *EMNLP 2010 — Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. MIT, MA, USA, October 9–11, 2010*, pp. 240–250.

Jing Wang, Mohit Bansal, Kevin Gimpel, Brian D. Ziebart & Clement T. Yu (2015): A Sense-Topic Model for Word Sense Induction with Unsupervised Data Enrichment. In: *Transactions of the Association for Computational Linguistics*, 3: 59–71.

Amy Beth Warriner, Victor Kuperman & Marc Brysbært (2013): Norms of valence, arousal, and dominance for 13,915 English lemmas. In: *Behavior Research Methods*, 45(4): 1191–1207.

Warren Weaver (1955): Translation. In: William N. Locke & A. Donald Booth (eds.) *Machine Translation of Languages: Fourteen Essays*, pp. 15–23. MIT Press and Wiley & Chapman. [Written 1949].

Julie Weeds, David Weir & Diana McCarthy (2004): Characterising Measures of Lexical Distributional Similarity. In: *COLING 2004 — Proceedings of the 20th International Conference on Computational Linguistics. Geneva, Switzerland, August 23–27, 2004*, pp. 1015–1021.

Laura Wendlandt, Jonathan K. Kummerfeld & Rada Mihalcea (2018): Factors Influencing the Surprising Instability of Word Embeddings. In: *NAACL-HLT 2018 — Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Long Papers. New Orleans, LA, USA, June 2–4, 2018*, pp. 2092–2102.

Joachim Wermter & Udo Hahn (2004): Collocation Extraction Based on Modifiability Statistics. In: *COLING 2004 — Proceedings of the 20th International Conference on Computational Linguistics. Geneva, Switzerland, August 23–27, 2004*, pp. 980–986.

Norbert Wiener (1955): *Cybernetics or Control and Communication in the Animal and the Machine*. MIT Press, 2nd edition. [Reprint, first published 1948].

Derry Tanti Wijaya & Reyyan Yeniterzi (2011): Understanding semantic change of words over centuries. In: *DETECT 2011 — Proceedings of the 2011 International Workshop on DETecting and Exploiting Cultural diversiTy on the Social Web @ CIKM 2011. Glasgow, U.K., October 24, 2011*, pp. 35–40.

Peter Wunderli (2013): *Ferdinand de Saussure: Cours de linguistique générale — Zweisprachige Ausgabe französisch-deutsch mit Einleitung, Anmerkungen und Kommentar*. Narr.

Yang Xu & Charles Kemp (2015): A Computational Evaluation of Two Laws of Semantic Change. In: *CogSci 2015 — Proceedings of the 37th Annual Meeting of the Cognitive Science Society. Pasadena, CA, USA, July 22–25, 2015*, pp. 2703–2708.

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao & Hui Xiong (2018): Dynamic Word Embeddings for Evolving Semantic Discovery. In: *WSDM 2018 — Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. Marina Del Rey, CA, USA, February 5–9, 2018*, pp. 673–681.

Tae-Won Yoon, Sung-Hyon Myaeng, Hyun-Wook Woo, Seung-Wook Lee & Sang-Bum Kim (2018): On Temporally Sensitive Word Embeddings for News Information Retrieval. In: *Proceedings of the NewsIR'18 Workshop @ ECIR 2018. Grenoble, France, March 26, 2018*, pp. 51–56.

Torsten Zesch & Iryna Gurevych (2006): Automatically creating datasets for measures of semantic relatedness. In: *Proceedings of the Workshop on Linguistic Distances @ COLING-ACL 2006. Sydney, Australia, 23 July 2006*, pp. 16–24.

Yating Zhang, Adam Jatowt, Sourav S. Bhowmick & Katsumi Tanaka (2015): Omnia Mutantur, Nihil Interit: Connecting Past with Present by Finding Corresponding Terms across Time. In: *ACL-IJCNLP 2015 — Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing: Short Papers. Beijing, China, July 26–31, 2015*, pp. 645–655.

Yating Zhang, Adam Jatowt, Sourav S. Bhowmick & Katsumi Tanaka (2016): The past is not a foreign country: Detecting semantically similar terms across time. In: *IEEE Transactions on Knowledge and Data Engineering*, 28(10): 2793–2807.

Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston & Bernhard Schölkopf (2003): Learning with Local and Global Consistency. In: *NIPS 2003 — Advances in Neural Information Processing Systems 16. Whistler, British Columbia, Canada, December 09–11, 2003*, pp. 321–328.

Theodore Ziolkowski (1969): Das Nachleben der Romantik in der modernen deutschen Literatur. Methodische Überlegungen. In: Wolfgang Paulsen (ed.) *Das Nachleben der Romantik in der modernen deutschen Literatur*, pp. 15–32. Lothar Stiehm Verlag.

Xiaojun Zou, Ni Sun, Hua Zhang & Junfeng Hu (2013): Diachronic corpus based word semantic variation and change mining. In: *IIS 2013 — Proceedings of the 20th International Conference Intelligent Information Systems. Warsaw, Poland, June 17–18, 2013*, pp. 145–150.