Routledge
Taylor & Francis Group

Historical Methods: A Journal of Quantitative and Interdisciplinary History

ISSN: 0161-5440 (Print) 1940-1906 (Online) Journal homepage: https://www.tandfonline.com/loi/vhim20

Digital begriffsgeschichte: Tracing semantic change using word embeddings

Melvin Wevers & Marijn Koolen

To cite this article: Melvin Wevers & Marijn Koolen (2020) Digital begriffsgeschichte: Tracing semantic change using word embeddings, Historical Methods: A Journal of Quantitative and Interdisciplinary History, 53:4, 226-243, DOI: 10.1080/01615440.2020.1760157

To link to this article: https://doi.org/10.1080/01615440.2020.1760157

HISTORICAL METHODS: A JOURNAL OF QUANTITATIVE AND INTERDISCIPLINARY HISTORY
2020, VOL. 53, NO. 4, 226–243
https://doi.org/10.1080/01615440.2020.1760157

Routledge
Taylor & Francis Group

Digital begriffsgeschichte: Tracing semantic change using word embeddings

Melvin Wevers[a,b] and Marijn Koolen[c]

[a]DHLab, KNAW Humanities Cluster, Amsterdam, The Netherlands; [b]Department of History, University of Amsterdam, Amsterdam, The Netherlands; [c]Digital Infrastructure, KNAW Humanities Cluster, Amsterdam, The Netherlands

ABSTRACT

Recently, the use of word embedding models (WEM) has received ample attention in the natural language processing community. These models can capture semantic information in large corpora of text by learning distributional properties of words, that is how often particular words appear in specific contexts. Scholars have pointed out the potential of WEMs for historical research. In particular, their ability to capture semantic change might assist historians studying conceptual change or specific discursive formations over time. Concurrently, others voiced their criticism and pointed out that WEMs require large amounts of training data, that they are challenging to evaluate, and they lack the specificity looked for by historians. The ability to examine semantic change resonates with the goals of historians such as Reinhart Koselleck, whose research focused on the formation of concepts and the transformation of semantic fields. However, word embeddings can only be used to study particular types of semantic change, and the model's use is dependent on the size, quality, and bias in training data. In this article, we examine what is required of historical data to produce reliable WEMs, and we describe the types of questions that can be answered using WEMs.

KEYWORDS

Conceptual history; word embeddings; digital history; semantic change

Introduction

With the large-scale digitization of historical sources in recent years, historians can now search through archives consisting of thousands of documents using keyword searches (Bingham 2010; Nicholson 2013). Such keyword searches can point us in the direction of particular uses of words, offering us a perspective on the formations of meaning expressed through the relationships between words in particular (historical) contexts. Simple frequency plots of single keywords, produced by, for example, n-gram viewers, offer no information on the relationships between words in historical contexts. To grasp how a particular word is situated historically and how relationships between words developed, historians need to turn to more sophisticated computational methods.

A particular method that received much attention in the Natural Language Processing (NLP) community, and more recently, in the field of Digital Humanities is Word Embedding. A Word Embedding Model (WEM) contains semantic and syntactic information, and it is constructed from the distribution of words in texts. The distribution of words refers to the frequency in which words co-occur with other words in an extensive collection of texts. WEMs are created by learning algorithms that extract relationships between words from large amounts of texts. These relationships can then be used to study the contexts of words.

The embedding of a word is a representation that is extracted from its context. For example, in the sentence "the quick brown fox jumped over the lazy dog", the word "fox" is surrounded by "the quick brown" and "jumped over the." For every single occurrence of every word in a corpus, the algorithm learns these contexts, of which the size, that is, the number of words on either side, can be arbitrarily selected. As a consequence, words that appear in similar contexts will also be similarly embedded. The embedding not only stores a word and its neighbors but also uses the information on its neighbors' neighbors to learn a word's context. This contextual information can be used for a wide range of lexical-semantic tasks, such as synonym detection, concept

CONTACT Melvin Wevers melvin.wevers@dh.huc.knaw.nl DHLab, KNAW Humanities Cluster, Amsterdam, The Netherlands

categorization, estimating semantic relatedness, and inferring analogous concept relations (Baroni, Dinu, and Kruszewski 2014).

In theory, word embeddings and the semantic tasks they enable offer historians the ability to study words in relation to other words, both synchronously and diachronically. Rather than searching for keywords, historians can search for the contextual *embedding* of words. For instance, searching with the keyword "democracy" retrieves only texts that explicitly mention the word "democracy," but searching with its contextual embedding also retrieves texts that describe the context in which the word "democracy" is used. In addition, historians can compare shifts in meaning over time, by, for example, comparing the embeddings of "democracy" in different text genres or discursive communities.

Scholars have highlighted the potential of WEMs for historical research, for the specific purposes of tracing conceptual change or studying discursive spaces (Azarbonyad et al. 2017; Hamilton, Leskovec, and Jurafsky 2016a; Kenter et al. 2015; Orlikowski, Hartung, and Cimiano 2018; Recchia et al. 2016). These approaches view the usefulness of word embeddings predominantly from the perspective of a computational linguist. Others have combined word embeddings with information on emotional valency of words (Hellrich, Buechel, and Hahn 2018). At the same time, computational linguists voiced their criticism and questioned the usefulness of WEMs (Dubossarsky, Weinshall, and Grossman 2017; Hellrich and Hahn 2016; Sommerauer and Fokkens, 2019).

With these concerns in mind, the central question of this paper is: how can word embeddings be used for historical research, and what are their limitations? We set out to answer this question by describing best practices for creating and evaluating WEMs, and we provide examples of the types of semantic change that can be studied. We argue that WEMs have their use, but for that to be the case, we need to take into account specific requirements in terms of data, and we need to be aware of what types of questions can be answered using this computational technique.

The critical voices have raised three crucial points. First, word embeddings require large amounts of training data to create reliable models, as well as substantial computing power. For many historians, the amount of digitized material is limited, and the data is often not directly accessible.

Second, even in the case of the availability of large-scale training data, the reliability of the produced models still needs to be assessed carefully. Assessing the quality of a model is notoriously tricky, as Section, "**Evaluating the quality of a trained word embedding model**" discusses. The quality depends on the size, the quality, and the possible bias in the data. In the case of historical data, imperfect Optical Character Recognition (OCR) introduces errors and increases the amount of spelling variation. A proper evaluation is then needed to assess whether shifts in word meanings captured in embeddings are reliable shifts and not caused by flawed OCR or too little training data.

Finally, most of the articles on word embeddings are evaluated using particular large-scale data sets, such as the Google News Corpus[1] or Corpus of Historical American English (COHA).[2] The semantic information captured in word embedding models is specific to the language used in the texts on which the models are trained. If the models are based on twenty-first-century American news articles, the word meanings reflect the use of that period, language, region, and genre. Historians often work with their own corpora that are, more often than not, much smaller than the corpora that researchers in computational linguistics use, and as a consequence, there is less contextual data for individual words to learn meaningful embeddings. These historical research corpora also tend to contain domain-specific language, as is the case with parliamentary data or scientific correspondences. While this domain-specificity can also help to study, for example, the possible bias between domains, domain-specific corpora are ostensibly much smaller than concatenated, heterogeneous corpora that represent more of the variation in natural language use.

While the concerns mentioned above are undoubtedly valid, the benefits and limitations of word embedding models are, in most of the cases, not discussed from the perspective of the historian. Taking the limitations into account, we argue that WEMs still have their use for historical research. At the same time, in digital history, there occasionally exists an almost unrealistic belief in the possibilities of word embeddings for historical research. With this article, we aim to offer a corrective and hope to provide a pragmatic, realistic view on the role of word embeddings for historical research.

The article is structured as follows, in Section, "**A short primer on word embedding**" we offer a short primer on word embedding and briefly explain how this method works. Second, we tie the possibilities offered by word embedding to theories on the

Table 1. Example co-occurrence matrix.

	Car	Motorcycle	Highway	Helmet	Cauliflower
Car	–	2	3	1	0
Motorcycle	2	–	3	3	0
Highway	3	3	–	3	0
Helmet	1	3	2	–	0
Cauliflower	0	0	0	0	–

historical study of concepts (Section, "**The link between conceptual history and word embedding**"). The work of, amongst others, Reinhart Koselleck has been instrumental in defining *begriffsgeschichte*, or conceptual history (Koselleck 1989). Even though Koselleck's work can be somewhat obtuse, it does offer a perspective on how historians can study semantic change as a proxy for broader historical processes. Third, we examine best practices for creating, evaluating, and using WEMs for historical research in Section, "**Considerations before training a model**" and Section, "**Evaluating the quality of a trained word embedding model**". In the former, we focus on the requirements in terms of data, while the latter deals with the evaluation of word embeddings. Fourth, in Section, "**Putting word embeddings to use**", we will give specific examples of how word embeddings can be used to examine semantic change and how they can also be used to facilitate search in historical archives. Finally, Section, "**Conclusion**" offers concluding remarks that can help historians decide whether and how they should consider using WEMs on data available to them. Moreover, we discuss the types of questions that can be answered using these models. Lastly, we highlight recent developments in Natural Language Processing, more specifically related to Deep Contextualized Word Embeddings, and their possible use for historical research.

A short primer on word embedding

In 1954, the linguist Zelig Harris demonstrated that semantically similar words share contexts (Harris 1954). He demonstrated that the distribution of contextual words— the vocabulary surrounding a particular word—defines the meaning of a word. For instance, in texts on cars and motorcycles, the words "car" and "motorcycle" are frequently surrounded by the words: "highway", "gasoline", and "speed." A word such as "helmet" is more common in the neighborhood of "motorcycle" than "car." Still, "helmet" is more closely related to "car" through its vicinity to the other surrounding words than it would be to, for example, "cauliflower."

The distributional hypothesis put forth by Harris forms the conceptual basis for word embedding. A word embedding is a semantic model that represents the co-occurrences of words in a multidimensional space. Each word w in a vocabulary V is represented as a continuous vector of a fixed dimensionality, or length, N. More concretely, a word's context is represented as a list of N numbers—a vector—that is learned from large amounts of text. The length of this vector is determined by the user. The resulting model can be represented as a matrix of size NxV. For every word in a vocabulary, we have a vector that represents that word's context.

Traditionally, a co-occurrence model consists of a matrix that represented the entire corpus (see Table 1). This matrix would include how often each word would appear in the context of every other word in the vocabulary. If "cauliflower" is the 537th word in the vocabulary and "helmet" the 963rd word, the 537th row in the matrix would contain the co-occurrence frequencies of "cauliflower" and all other words in the vocabulary, with the 963rd column in that row being the co-occurrence of "cauliflower" and "helmet." With this matrix, a word like "helmet" can be represented not only as just the word "helmet" but also as a context vector that describes it as somewhat related to "car," but more related to "motorcycle" and "highway." Historians can learn about the contexts of words by exploring co-occurrences of specific keywords using tools such as AntConc.[3]

Early techniques that model word associations as geometric relationships were used in e.g., Information Retrieval, in which the associations between words is calculated using the co-occurrence of words (Giuliano and Jones 1962). A more sophisticated technique was developed by Salton et al. (Salton, Wong, and Yang 1975) in which documents are represented as vectors in a Vector Space Model, where every word in the vocabulary is a dimension in this high-dimensional space. Documents are represented in a term-document matrix as vectors, where the frequency of a word in that document determines its extension in that word's dimension. Angles and distances between documents and words can be calculated to find documents relevant to a set of query keywords.

However, for large corpora, consisting of many hundreds or thousands of distinct words, co-occurrence matrices could turn out to be incredibly expansive, making them computationally intensive to create and to use. At the same time, these large matrices are also sparse; that is, most of the values in the matrix are zero. In Table 1, the word "cauliflower" does not co-occur with any of the other words. However, the same is true for most words in the vocabulary, as they

are relatively infrequent in the corpus, and thus seldom appear in the context of other words. This sparsity impacts mathematical operations as they have to extract word relationships from *weaker* patterns of information. A further development was Latent Semantic Indexing (Deerwester et al. 1990), which transforms the high-dimensional vector space to a space with far fewer dimensions by exploiting higher-order structure in the term-document matrix. The higher-order structure reveals itself through those same statistical associations between words that often cooccur. The dimensions of frequently cooccurring words are correlated. By reducing the number of dimensions using these statistical associations, a latent semantic space is created, in which documents can be associated with words even if those words do not occur in those documents.

Modern algorithms, such as Word2Vec, offer a faster, more efficient way of constructing these matrices as embedding spaces (Mikolov, Sutskever, et al. 2013).[4] This algorithm, which is a learning system that relies on neural network technologies, constructs word representations one sentence at a time. For the algorithm to be able to create a reliable word embedding, researchers need to input large amounts of data, containing many sentences, from which the algorithm *learns* the contexts between words.

The resulting word context is represented in a predefined number of dimensions (N). As a consequence, not every single word's relation to every other word is explicitly represented. Instead, the algorithm learns a "compressed" representation of the context within a fixed dimensionality. This process produces a dense matrix—as opposed to a sparse matrix—of, for example, 300 dimensions. The consequence of this compression is that the resulting dimensions represent implicit meaning, which is why dimensions are sometimes referred to as *latent factors*. The same implicit meaning is present in the co-occurrence matrix where each word represents a dimension, but that meaning is hidden or latent across many different dimensions. Dimension reduction techniques squash, twist, and warp the matrix so that the resulting 300 dimensions more directly represent those implicit meanings.

Even though semantic and syntactic regularities are preserved in the compressed word embedding models, one does not know beforehand what linguistic information is represented in a particular dimension (Mikolov, Yih, et al. 2013). Some dimensions might capture semantic information, while others capture syntactic information. A dimension could, for instance, capture superlatives. Words such as "greater," "smaller," and "better" all appear close to each other in that dimension.

Learning word embeddings

The learning process typically uses a window of a fixed number of words around each word, e.g., two words before and after, and slides it over the sentence to extract the context of each word. The size of the window span— size of the context—is one of the parameters that the user can set before training an embedding model. Longer window spans tend to contain more semantic information, while shorter preserve syntactical information. This example shows the influence of training parameters when constructing a WEM. There is no single optimal parameter setting in most cases; decisions regarding parameters are linked to the task and data at hand, as well as the underlying research question. Experience in training models can help to improve the decision-making process.

There are two strategies of learning word representations in an embedding space: Skip-Gram Negative Sampling (SGNS) and Continuous Bag-of-Words (CBOW). The former, SGNS, predicts contexts—words within a set window span—for a given word, while CBOW does the reverse and predicts words from a given context (see Figure 1). Turning back to our earlier example, for the sentence "the quick brown fox jumped over the lazy dog" and a window size of one word on each side, SGNS would predict "brown" and "jumped" based on the input word "fox." CBOW, on the other hand, would predict "fox" given the words "brown" and "jumped." The skip-gram method is commonly preferred for semantic tasks (Levy, Goldberg, and Dagan 2015). However, there is some debate as to which methods perform better for uncommon words in a corpus (Mikolov, Sutskever, et al. 2013; Sahlgren and Lenci 2016).

Geometric operations in the word embedding space

A particularly useful feature of Word Embedding is that semantic and syntactic information is related to distances and directions in the embedding space (Baroni, Dinu, and Kruszewski 2014; Miller and Charles 1991). The most well-known aspect is that the semantic similarity between two words is represented by the closeness in the embedding space of the two vectors related to those words. Closeness is represented using cosine distance, which indicates the angle between two vectors along two or more dimensions or latent factors (see

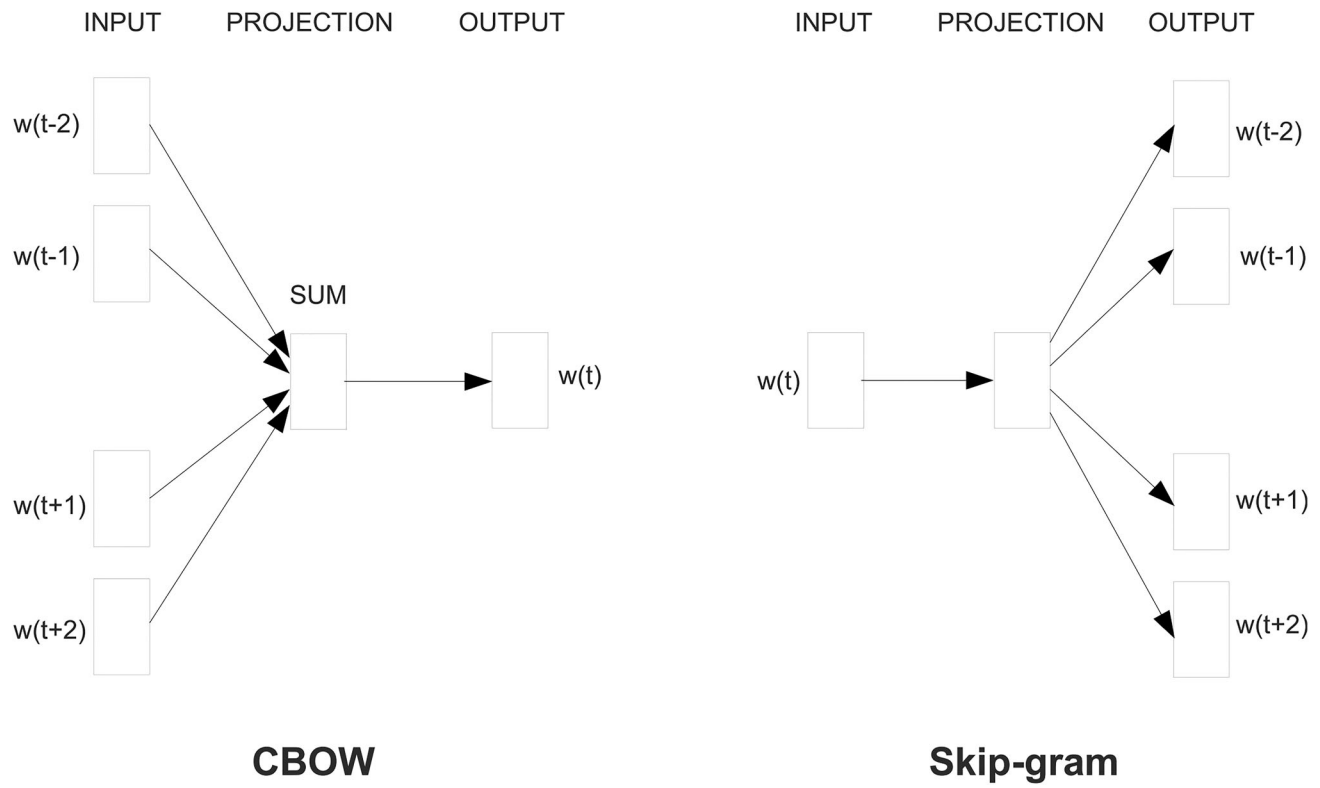

Figure 1. Continuous bag-of-words and skip-gram architectures (Figure taken from: (Mikolov, Sutskever, et al. 2013)).

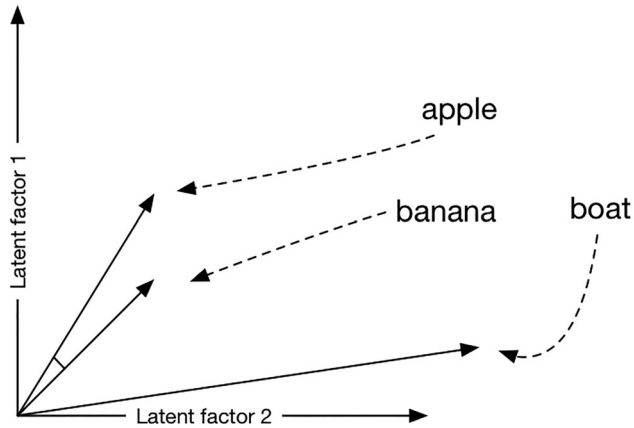

Figure 2. The angle or cosine distance between words along two dimensions (latent factors) indicates how close they are in meaning.

Figure 2). A smaller angle indicates semantically similar words; in this case, "apple" and "banana." This feature of word embeddings enables researchers to ask for similar words by querying the nearest neighbors, as represented by a relatively small cosine distance, for a particular word or group of words. For example, querying for similar words to "apple" returns the vectors of words related to other fruits.

The relation between word meaning and the geometry of the embedding space goes much further than nearest neighbors. Mikolov et al. demonstrated the possibility of finding capital cities by querying country names (Mikolov, Sutskever, et al. 2013). A movement of a given distance in a particular direction has a relatively stable meaning. For instance, the movement required to go from the word "France" to the word "Paris," is almost the same as that required to go from "Germany" to "Berlin" or from "Russia" to "Moscow." In other words, that specific movement translates to "X has capital city Y." Figure 3 shows how this particular semantic feature is captured in the embedding space. It also demonstrates that related places, such as "Athens" and "Rome" are closer to each other in this representation. To find these semantic consistencies in the data, we need to create specific ways of querying the embedding space. From the produced output, we can learn that our question is relevant, expressed by the consistent relationship between countries and capitals. Besides, we learn that there is a structure between these relationships.

The relation between meaning and geometrical representation means that we can apply linear algebra to examine the embedding space. This allows researchers to ask more complicated questions such as: what is to C as A is to B (Levy and Goldberg 2014)? The canonical example here is: "king" is to "queen" as "man" is to what? In this instance, the answer, of course, is "woman." The following equation expresses this question:

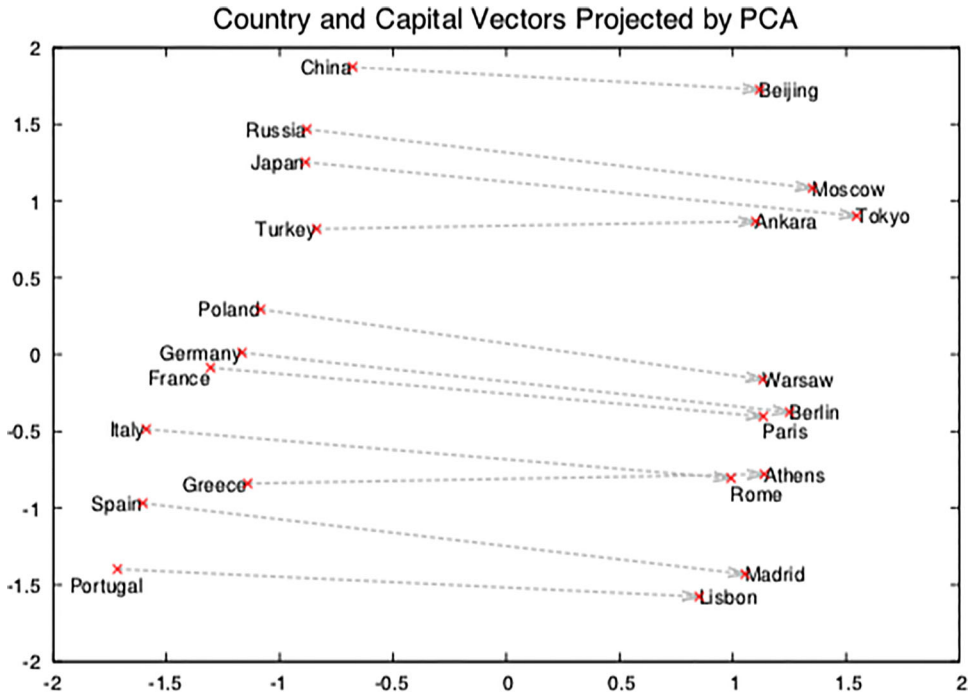

Figure 3. Two-dimensional projection of vectors of countries and their capital cities. (Figure taken from: (Mikolov, Sutskever, et al. 2013).

$$\text{Vking} - \text{Vman} + \text{Vqueen} = \text{Vwoman} \quad (1)$$

Semantic change can also be traced by comparing the cosine distance between word vectors in different periods. Researchers can compare vectors—semantic contexts—for the same sets of words in different periods (Hamilton, Leskovec, and Jurafsky 2016a, 2016b). The same holds here: greater distances refer to more considerable semantic changes. Popular examples include the semantic shift of words such as "gay" and "cell." Where "gay" used to refer to a cheery disposition, it is now commonly used to indicate sexual orientation. Similarly, "cell" was shorthand for a prison cell, while its dominant use currently concerns cellular phones.

These two examples are cases in which the meaning of a word changed almost diametrically. More often than not, discursive shifts are more subtle and might better be represented by changes in the network of relationships between particular words (Kenter et al. 2015). It might be the case that most words in a semantic context stay the same, but that the disappearance or appearance of one word is a clear sign of semantic changes, even though the general meaning of a word stays relatively stable.

This section provided a brief overview of the theoretical underpinning of word embedding models and how they are constructed.[5] We will now turn to the perspective of the conceptual historians, which we argue can be connected to the notion of word embeddings. These connections offer a framework that helps to discern what types of historical questions might be answered using word embeddings.

The link between conceptual history and word embedding

Studying discursive and semantic changes as indicators of broader cultural-historical transformations is an elemental part of the study of history (Koselleck 1989). Historians, such as Reinhart Koselleck, Michel Foucault, and Quentin Skinner, have examined why particular meanings took up prominence while others dissipated (Foucault 2012; Koselleck 2002; Skinner 2002). The subfields of conceptual history and historical semantics focus specifically on semantics from a historical perspective (Allan and Robinson 2012; Junge and Postoutenko 2014; Koselleck 2002). By looking into linguistic changes, historians have tried to uncover the codifiers and shapers of particular meanings. Koselleck points out that not only the changes in semantics but also how words and their associated meanings change shapes our political and social experience. Such genealogies of meaning can shed light on cultural-historical developments, and these genealogies can also inform our contemporary views on society (James and Steger 2014).

Terms and their meaning often point toward the notion of a concept—an abstracted idea or generalized

notion. The *concept* of a concept, however, is highly debated. Of interest for this article is the distinction made by scholars working in the German tradition of "Begriffsgeschichte" and Anglo-Saxon school of "the History of Ideas" (Brunner, Conze, and Koselleck 2004; Richter 1987; Schmieder 2019). The major difference between the History of Ideas and *Begriffsgeschichte*, Richter notes, is that the former works with the notion of unchanging "unit-ideas" that occurred throughout history. The latter employs linguistic principles to study the continuities and changes in the use of concepts. Consequently, it is more closely related to historical semantics, while the History of Ideas resonated more clearly with philosophy. An in-depth discussion of these schools reaches beyond the scope of this article. For now, it suffices to note that Koselleck's views on *Begriffsgeschichte* and conceptual history reverberate with elements of distributional semantics and the possibilities offered by word embeddings.

The question that lingers is: how do conceptual historians, such as Koselleck, approach the concept of a "concept"? Reinhart Koselleck views a "concept" as a word that contains a wide range of social and political meanings and connotations.[6] As such, concepts turn raw experience (*Erfahrung*) into lived experience (*Erlebenis*) (Palti 2011). Because of its inherent ambiguity, a concept functions as a space of signification in which meaning is contested. This ambiguity makes a structural analysis of concepts notably difficult and opaque (Andersen 2003).

In Koselleck's work on concepts, he displays a particular interest in the "onomasiological" relation between words. This particular branch of linguistics is invested in terms that represent particular concepts. For example, one can refer to concepts such as "democracy" or "equality," through the keywords "democracy" and "equality." However, these concepts represent more than just these indexical words, and for this reason, networks of words or discursive spaces might better represent concepts. In other words, how are words used and in which contexts are words used might tell us about the meaning of these concepts in particular historical times. This notion reverberates with ideas by Ludwig Wittgenstein, who wrote that "meaning of a word is its use in the language." He argued that by tracing the contextual shifts of words, we could "travel with the word's uses through a complicated network of similarities overlapping and criss-crossing" (Wittgenstein 2010, p. 66). Such networks of words reveal the "architecture of concepts—the words, phrases, sentences, and statements that we use and use us (De Bolla 2013)."

Koselleck's central aim was to examine the formation of concepts and the transformations of semantic fields. The relationships between concepts make up semantic fields. Linguists Dan Jurafsky and James Martin offer the following definition of a semantic field: "a set of words which cover a particular semantic domain and bear structured relations with each other (Jurafsky and Martin 2009)." Semantic fields can be studied diachronically, or synchronously. A diachronic analysis deals with the origin and transformation of specific concepts. Put differently, when did relationships between words form, and how did they change over time? The latter deals with the analysis of the semantic field in which concepts appear and their connection with other concepts (Andersen 2003).

A concept, thus, acquires its meaning from its use in specific historical contexts, represented by its surrounding words, as well as the relationship these surrounding words have in opposition to other groups of words (Skinner 2002). There is no clear demarcation between semantic fields, but opposing structures imbue concepts with meaning. The challenge is then how to find such structures, and can word embedding possibly help in studying them?

Connection to word embedding

In a word embedding, a single word is represented by a vector. In practice, this means that we can use the vector for "apple" as input and ask for similar vectors. However, there are different degrees of similarity. Put differently, similar words can appear in different semantic fields, in which they are imbued with different meanings. In the case of "apple," the word is used in the context of fruits, but also as the name of a technology company, which places the word in the contexts of other companies, such as IBM, Google, and Microsoft.

Words and their specific networks can be isolated through algebraic operations on vectors, for example, adding vectors together, representing particular concepts in the semantic space. This method can help to pinpoint the use of words in specific semantic fields. The word "plane" has an ambiguous meaning, indicating both a flat surface and an airplane. By adding word vectors to the vector of "plane" that unequivocally point to one of the two meaning we can search for the context in that particular area in the semantic space. The summation of *pilot* + *plane* + *landing* points in the direction of the meaning of airplane,

while summing *line* + *plane* + *angle* more clearly signals the use of plane as a flat surface (Gavin 2015).

These summed vectors can help to isolate distinct discursive spaces, although there is always a degree of overlap between these spaces. Therefore, constructing these summed vectors requires domain knowledge and is dependent on the corpus. Researchers always need to check how the summed vectors are represented in the embedding space before they can be used meaningfully to query a distinct discursive space.

This process of summing vectors resonates with ideas that Koselleck voiced related to semantic fields. The concept of "state" can, for instance, be represented through words such as "sovereignty" and "territory." These words and their relationships to neighboring concepts determines the meaning of the concept "state." Put differently, Koselleck argued that semantic fields could be delineated by juxtaposing them to fields containing antonyms, representative of counter-concepts (Ifversen 2003). Researchers can compare the similarities between word vectors, or groups of vectors, based on their cosine distance. The ability to also juxtapose different clusters of words shares similarities with Koselleck's *Begriffsgeschichte*, albeit that his approach mostly focuses on discursive developments over extended time periods.

This connection between (summed) vectors and concepts can also be used to do a full-text search in historical corpora with more than just keywords. In Section, "**Word embedding and search in historical corpora**" we detail how, for a search query with multiple keywords, the sum of their vectors can be used to disambiguate these words (e.g., "state" as "sovereignty" or as "condition"). We also describe how we can expand the query with words close to the summed vector to find documents that discuss the same concepts but that do not mention the query keywords. In Section, "**Word embeddings to trace semantic change**", we demonstrate how word embeddings offer a way to study semantic continuity and change in a synchronic and diachronic fashion.

Considerations before training a model

This section discusses what is required in terms of data to be able to train word embedding models that can be helpful for historians. For historians who consider using word embeddings in their research, we recommend taking into account at least three central aspects of the data when training word embeddings: corpus size, OCR quality and spelling variation, and bias.

Size of the data set

First, the size of the training data is the most significant determinant of the quality of the word embedding (Tulkens et al. 2016). For robust representations, algorithms, such as Word2Vec require texts that contain words that frequently co-occur. If there is not enough data, the word embedding algorithm will not be able to *learn* word representations accurately. The Google News Corpus, for example, contains over 100 billion words. However, how do we determine what the minimum corpus size should be?

Hamilton et al. have argued that a reliable model requires at least a corpus of 100 million words per time slice and a vocabulary of around one to two million distinct words. For smaller corpora, they advise using co-occurrence matrices rather than word embeddings. An extra complication is that the heterogeneity of the corpus. A corpus of texts from a single narrow domain and genre probably has a smaller vocabulary and more consistent contexts than an equally sized corpus of texts from many different domains and genres.

In the case of historical data, researchers often do not have access to large, diachronic sets of digitized textual data. Data sets that are regularly used to study semantic change include parliamentary data, newspapers, books, and journals. Unfortunately, this data has been digitized for only a few languages and often not for extended periods. Another option to increase the amount of data is to combine different types of sources. However, combining corpora does not always improve the quality of the embeddings (Tulkens et al. 2016).

Additionally, when working with changes over time, we need to work with multiple word embedding models that capture word use for different periods. To be able to produce multiple models, we need to slice the data into subsets, effectively decreasing the number of words per model, and thus reducing its reliability. Again, there is no golden rule for how to approach this issue. In part, it is related to what type of question one wants to answer, and second, to the resulting quality of the data. In Section, "**Evaluating the quality of a trained word embedding model**", we discuss several methods to evaluate whether the data is reliable.

Moreover, one should be careful when training models on data from different periods. Before one can measure word similarities between different embeddings, we first need to align the embedding spaces. Models do not align when trained independently, meaning that areas in one semantic space are not

necessarily in the same space for a different model. Even though the local structure of two embedding models might be similar—that is, the same words are near to each other—the global structure might differ. In this case, clusters of semantically similar words might be positioned in different locations within the embedding space. In order to meaningfully compare cosine distances, we first need to align the embedding spaces.[7] Plainly put, this can be compared to making sure that the north is facing upwards when comparing maps made in different periods.

A way to align models is to use post-training alignment (Hamilton, Leskovec, and Jurafsky 2016a). Alignment algorithms attempt to reshape the semantic space, ensuring that multiple spaces are globally aligned. One such technique, Procrustes Alignment, adapts all models to the first model to make sure that the spaces are globally aligned. A downside of this method is that it requires the vocabularies—the sets of distinct words—in the embedding spaces to be equivalent (Hamilton, Leskovec, and Jurafsky 2016c). Consequently, words that are not present in all embedding spaces have to be pruned.

For historians interested in changes in semantics, this pruning step poses a problem. In language, new words appear, and old words disappear all the time, making diachronic alignment difficult. In the case of synchronic alignment of models based on different language genres or communities, the pruning misses the importance of differences in vocabularies. It makes the models seem more similar than they are. More advanced methods for alignment have been proposed (Barranco et al. 2018; Rudolph and Blei 2018; Yao et al. 2018). These, however, are more computationally intensive and are more challenging to implement.

Yoon Kim et al. propose a different solution, which entails dividing the data into chronologically-ordered bins and subsequently training a model on the first temporal sequence (Kim et al. 2014). Next, the initialized embedding space is used as input for the second model, which continues training on this model. In simpler terms, the model of a previous time slice is used as the point of departure for training a new model on the next time slice. When doing this, the model must be adaptive enough to let new information transform learned in formation from the previous model, effectively overwriting obsolete associations and words.

Another approach is to use overlapping windows, which also forces the algorithm to build upon earlier models (Kenter et al. 2015). This approach requires training bins consisting of, for example, five-year periods, where we shift the position only one year during training. For instance, 1950-1954 is followed by 1951-1955, then 1952-1956, and so on and so forth. Again, there is no one-size-fits-all solution. Depending on the data and the question, researchers need to choose an appropriate alignment method.

Ocr quality and spelling variation

The second aspect that needs to be taken into account is the amount of spelling variation of words, either because of OCR errors in the data or because of contemporary spelling variation or historical changes in spelling. Mistakes caused by flawed optical character recognition lead to an increase in word variation in the data. Because of OCR errors, the number of distinct word forms increases, even though many of these are not actual words. At the same time, their frequencies and, therefore, the number of contexts in which they appear, decrease. One way to account for this is by excluding infrequent words, which also removes many word variations resulting from faulty OCR when training the embedding. One needs to determine the threshold by trial and error since this threshold is very dependent on the corpus and the quality of the OCR.

One can also check whether words exist in a dictionary before adding them to the embedding model. This technique excludes variations due to OCR since these variations do not appear in the dictionary. Of course, this does not solve the problem of OCR errors, producing word variations that appear in the dictionary. This is an issue, especially for short words, where a difference of one character could change a word into a different word. For example, the word “bear” could turn into “beer.”

For spelling variation that is not caused by OCR errors, but inherent to language, there are various methods to normalize the spelling (Bollmann and Søgaard 2016; Piotrowski 2012; Richter et al. 2018). For instance, variation in historic spelling patterns of syllables can be detected statistically and normalized using so-called rewrite rules that change “hys” into ‘his’ so that “hystory” becomes “history” (Koolen et al. 2006). Of course, such rules are never perfect and occasionally change proper words into incorrectly-spelled words.

Yet another approach is to fix the word variations after training the embedding model. If the corpus is large enough, spelling variants of words will be in each other’s proximity in the embedding space, as they appear in very similar contexts. This closeness

allows researchers to spot common OCR mistakes and spelling variants and to group words that differ only slightly (Martinez-Ortiz et al. 2016). Still, the existence of these variants affects learning relationships between words. Above all, we need to be aware that OCR errors can significantly reduce word frequencies and, thereby, the number of usable words in a corpus, condensing a seemingly large corpus into a much smaller one.

Bias in the data

The final aspect that impacts the usability of the word embedding is the possible bias in the data. Bias, in this case, alludes to particular cultural or political perspectives present in the data that will subsequently become encoded in the word embedding. Given that they are strong enough, semantic structures that express certain ideas related to gender and ethnicity in the training corpus are also captured in the resulting embedding spaces. There have been attempts to remove these forms of bias from models (Bolukbasi et al. 2016), while others have used it as an indicator of historical situatedness of language (Azarbonyad et al. 2017; Garg et al. 2018).

These opposing views toward bias signpost a more substantial issue related to the use of word embeddings for historical research. Often word embeddings are used for tasks related specifically to Natural Language Processing. In this case, a large, comprehensive corpus yields an embedding that performs well on a wide array of tasks. In other words, this embedding model captures "natural language." However, historians are often interested in domain-specific language that is culturally, geographically, and historically situated. In this case, the bias in an embedding is a *feature* and not a bug. The challenge, however, is that such domain-specific corpora generally are relatively small, making them less useful to produce reliable word embeddings. Combining multiple domain-specific corpora can lead to specific word representations related to a particular domain to be overshadowed by more dominant representations in the corpus as a whole. In Section, "**Putting word embeddings to use**", we give an example of how word embeddings trained on different newspapers can be used to examine bias as expressed through semantic variations.

Evaluating the quality of a trained word embedding model

Researchers can turn to a set of evaluation methods to determine the quality of word embeddings models. In this section, we will describe how these evaluation methods work. There are two main methods of evaluation word embeddings: intrinsic and extrinsic evaluations (Wang et al. 2019). Intrinsic evaluations check the quality of the word representations by measuring syntactic and semantic relationships among words. Extrinsic evaluations use word embeddings as input for different Natural Language Processing tasks, such as part-of-speech tagging or named-entity recognition.

We recommend using both types of evaluation, as they are not only complementary but they can also shed light on each other. For the historical domain, one could also think of tasks related to answering a particular type of historical question (Sommerauer and Fokkens 2018). However, this type of evaluation is highly subjective and cannot be measured in absolute terms, but only through comparing different models.

Typical intrinsic evaluation tasks include 1) word similarity, 2) word analogies, and 3) overall model coherence measures. The first task uses predefined lists of similar word pairs. Ideally, the words in these pairs should be located close to each other in the embedding space. Proximity is thus an indicator of the model's quality (Batchkarov et al. 2016). However, it is notably difficult for annotators to score the similarity of words.

Word similarity can be expressed in two basic forms: *syntagmatic* and *paradigmatic* associations (Matsuoka and Lepage 2014; Rapp 2003). The former refers to words that frequently co-occur, which also have different grammatical roles. One can think of words such as "car" and "drive" or "judge" and "laws." Paradigmatic associations refer to words that can replace each other in a sentence without changing the grammatical structure of the sentence; for example, "speak" or "talk."

How a word embedding model is trained affects the representations of particular types of word similarity. The window size determines how many words that surround a target word—the context—are taken into consideration. A model trained using a small window size is more likely to capture syntagmatic relationships, since these words often occur in close proximity. Larger context windows allow for words that have a paradigmatic relationship, that is words with similar neighbors (Schütze and Pedersen 1993). The relationship between window size and type of word association is also language-dependent. In English, for instance, semantically related words are often grouped closer together than in Dutch. For this reason, to better capture paradigmatic associations,

larger window sizes are preferred when working with Dutch texts.

While word similarity tasks are a commonly-used evaluation method, it is not as straight-forward as it seems. Moreover, in the case of historical texts, we also have to take into account that similarities between words can change over time. An anachronistic list of word pairs might, therefore, not correctly assess the quality of historical models.

The second task, word analogies, is quite similar to the first. Rather than using similar words, evaluation relies on words that are analogous semantically or syntactically. Popular analogies include relationships between capitals and countries ("Paris" is to "France" as "Berlin" is to?) , verb inflections ("swim" is to "swam" as "go" is to?) , or nouns and adjective relationships.[8]

For historical data sets, these analogies are unreliable for three principal reasons.

First, many of the existing sets of analogies are in English and trained on contemporary English. Historians regularly work with different languages from diverse historical periods. Tulkens et al. created an analogy set for Dutch word embeddings (Tulkens et al. 2016). To be able to rely on analogy sets for evaluation purposes, we need to construct language and time-aware analogy sets. Second, analogies are often constrained to particular domains and historical periods. For example, the analogy: *Russia: Russian: : Ukraine: Ukrainian*, contains geographical information that is correct for contemporaneous data. These countries and these relationships are not represented in historical data. Moreover, there could also be analogies represented in data that are not part of the common analogy sets, giving the impression that the trained embedding model is under-performing. Third, historians are interested in semantic relationships and possible shifts, while these analogy lists assume a static relationality.

The third evaluation task focuses on establishing the coherence between different models. This is done by comparing a limited set of the nearest neighbors for every single word in multiple models. If there is a substantial degree of intersection between these words, this indicates a more reliable model (Hellrich and Hahn 2016). This method can be useful for comparing the effects of training parameters. Still, judgments based on word neighborhoods in trained models need to be approached with caution, as Hellrich and Hahn point out. They show that factors such as neighborhood size, word frequency, and word ambiguity impact the reliability of word neighborhoods. Although they present techniques to mitigate these issues—often with rising computational costs—they caution against relying too strongly on embeddings models. Sommerauer and Fokkens advise relying on control words and verifiable hypotheses that can help to establish whether models behaved as expected (Sommerauer and Fokkens 2019). Others have proposed sampling and shuffling of data to minimize random effects while training the model (Antoniak and Mimno 2018; Dubossarsky, Weinshall, and Grossman 2017). This sampling method, however, requires sizeable amounts of data; otherwise, the samples will become too small.

This last point ties in with another difficulty associated with WEMs for historical research, namely the lack of large-scale digitized resources. Even in the case of large corpora that produce WEMs that capture semantic information, this information is more reliable for highly frequent words. For uncommon words, there is often not enough information in the data set to determine an embedding. The considerations offered in this and the previous section can help to assess the reliability and use of the word embedding model for historical inquiry.

Putting word embeddings to use

In this section, we discuss examples to show how word embeddings can be used to trace semantic change and for purposes of search in historical archives.

Word embeddings to trace semantic change

There are several ways to examine semantic change using word embedding models, and we will highlight three of them.[9] The first method looks for semantic change on the word-level, measured by examining the global shift in a word's position between embedding spaces. To calculate this, we take a word's vector in two different periods and measure the cosine distance between them. Did the word itself change position?

In Figure 4, we see the shift between five-year periods of the words "abortus" (abortion) and "democratie" (democracy). For both words, the frequency increases over time—represented by the dashed line, while the position of abortion changes considerably while the semantic field of democracy stays relatively stable. This is a straightforward way of assessing whether words change semantically over time. This method can help to find points of inflection, or to draw comparisons between different

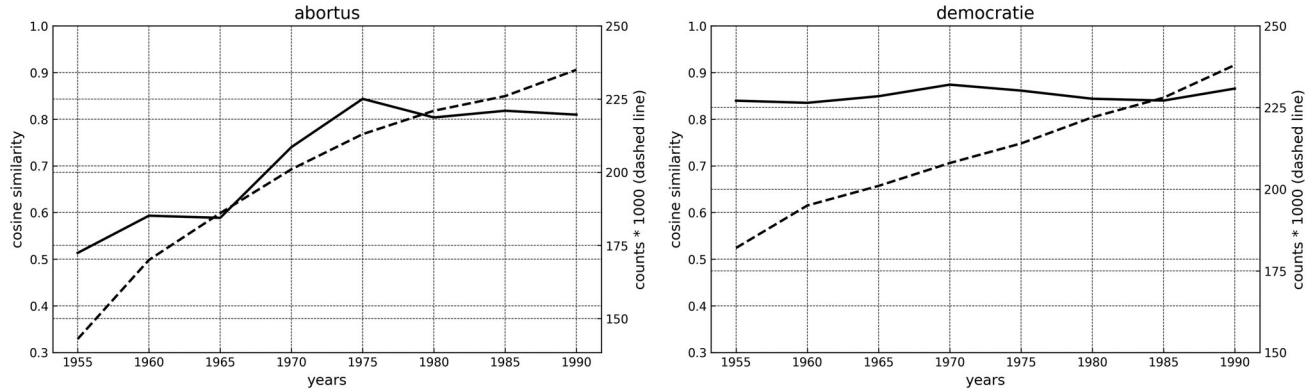

Figure 4. Semantic Shifts of Individual Words in Dutch Newspapers (dashed line indicates the frequency and the solid line refers to cosine similarity.

corpora, for instance, are certain semantic shifts to certain words more prominent in documents with a specific ideological signature?

The second method focuses on changes in a word's nearest neighbors—its most similar words. These local changes are found by finding the set of nearest neighbors of a target word across embeddings. Next, a similarity vector is calculated between the target word and every single word in the set of neighbors. Finally, this similarity vector is compared between different periods. Rather than measuring the position of a single word, we are capturing a target word's relation to its local neighbors. Comparing a word's similarity with its neighbors is indicative of local changes. Local neighborhood measures are more likely to pick up changes in the use of nouns, which are often indicative of cultural transformations. These measurements have been described as picking up semantic changes on the conceptual level (Hamilton, Leskovec, and Jurafsky 2016b).

Using the same target words as in the first example, we see, for instance, that the nearest neighbors of "abortion" change from words such as related to diseases such as "tuberculosis" and the generic "pregnancy" to more specific words such as "contraception", "sterilization", and "legalization" (see also Figure 5). From this information, we can already assess that the local context changes. Figure 6 shows how the local neighborhood changed more clearly for "abortion" than for "democracy." Closer inspection of the nearest neighbors can give insights into how the local changed.

Translations by authors, some Dutch synonyms do not exist in English and were thus combined. "Vruchtafdrijving" is a euphemistic term for "abortion" that cannot be translated.

In Section, "**The link between conceptual history and word embedding**", we discussed how we could also add and subtract vectors from each other to examine specific semantic fields. In what follows, we will highlight two articles that apply such methods to historical data. First, historians van Lange and Futselaar use WEMs trained on Dutch parliamentary debates to show discursive changes related to war criminals (van Lange and Futselaar 2018).[10] In their paper, they rely on summed vectors of words related to concepts such as "oorlogsmisdadiger" (war criminal). Next, they contrast the top 250 words most similar to "oorlogsmisdadiger" to words related to concepts such as "slachtoffer" (victim) and "verrader" (traitor). When doing this for different time periods they can show the relative distance between the concept "oorlogsmisdadiger" and the concept "slachtoffer" and "verrader", showing that "the war criminal vocabulary shifted from focusing on the act of crime committed by war criminals towards the consequences of these deeds for victims and relatives."

Wevers has trained multiple word embeddings models for different periods and newspapers with varying ideological backgrounds to investigate how gender was represented historically (Wevers 2019). This study builds upon the work of Garg et al. that also investigated gender bias through word embedding. Wevers extends this method by incorporating multiple historical sources, rather than using a comprehensive gold standard data set. The study constructs specific vectors using external lexicons, such as the Meertens database of Dutch first names[11], and HISCO (Historical International Classification of Occupations).[12] The gender vector, for example, is constructed by combining words such as "he", "father", "man" with popular male first names. This shows that the approach to constructing vectors and isolating semantic fields is not straightforward and benefits from domain knowledge and additional lexicons.

For six different newspapers, the study calculates the distance between target words, for example,

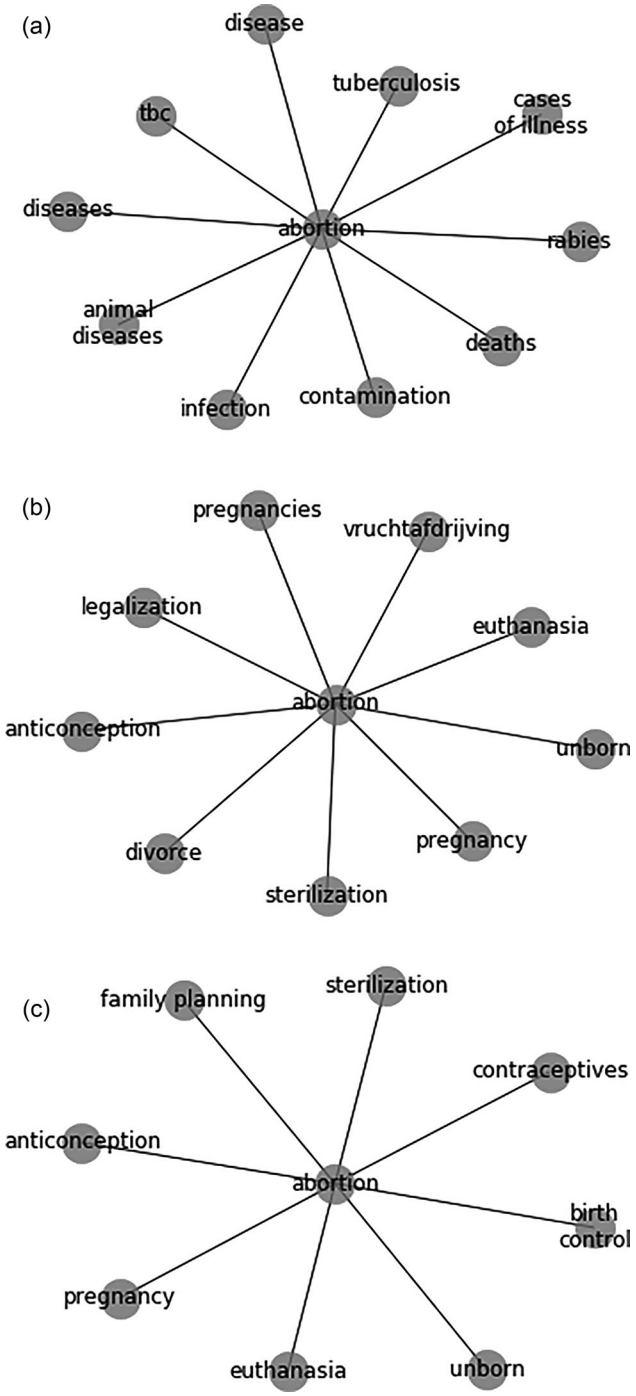

Figure 5. Nearest neighbors to "abortion" in 1950-1954, 1975-1979, and 1990-1994.

occupations and a summed vector that denotes gender, either male or female. The study concludes that individual newspapers show clear divergences in their biases and in the ways these biases change. We see that the newspapers with a social-democratic (*Vrije Volk*) and religious background, either Catholic (*Volkskrant*) or Protestant (*Trouw*), demonstrate the most evident shift in bias toward women. The liberal/conservative newspapers *Telegraaf*, *NRC Handelsblad*, and *Parool*, on the contrary, orient themselves more clearly toward men.

Word embedding and search in historical corpora

In addition to tracing semantic changes, word embeddings can be used in the context of search in digital archives and libraries, specifically for information retrieval and query expansion purposes.

A typical problem when searching with user-provided keywords is the so-called *semantic gap* between those keywords and the concepts or topics they represent, such that some relevant documents are not found because they describe the same concepts or topics with different words. There are techniques that search engines can employ to bridge or reduce this semantic gap by expanding the query with additional keywords, either automatically or interactively, by letting the searcher choose from a list of suggested keywords (Carpineto and Romano 2012; Efthimiadis 1996). Automatic query expansion, also called *blind relevance feedback*, retrieves documents using the query, then extracts significant terms from the highest-ranked documents, adds them to the query, and retrieves a new set of documents.

As mentioned at the end of Section, "**The link between conceptual history and word embedding**", WEMs can be an alternative source for query expansion and have certain advantages over the method described above (Diaz, Mitra, and Craswell 2016; Roy et al. 2016). The query is expanded with semantically related keywords by computing neighboring terms in the embedding space of the query keywords. The main advantage of this method over using blind relevance feedback is that the word embeddings are based on the entire corpus, instead of only the highest-ranked documents based on the initial query, using potentially more nuanced aspects of semantic similarity.

When the WEM is trained on different temporal subsets of the collections to capture semantic shifts and vocabulary shifts, it can also help searchers find documents that use older vocabulary even if they use modern language keywords. Huistra and Mellink argue that when historians select sources from digital archives, they should construct complex queries with multiple terms related to their search topic to deal with linguistic variation (Huistra and Mellink 2016). In such cases, WEMs can offer data-driven ways to suggest additional search terms for a given query. (Szymanski 2017) used WEMs to search for temporal word analogies. E.g. what phrase in documents from

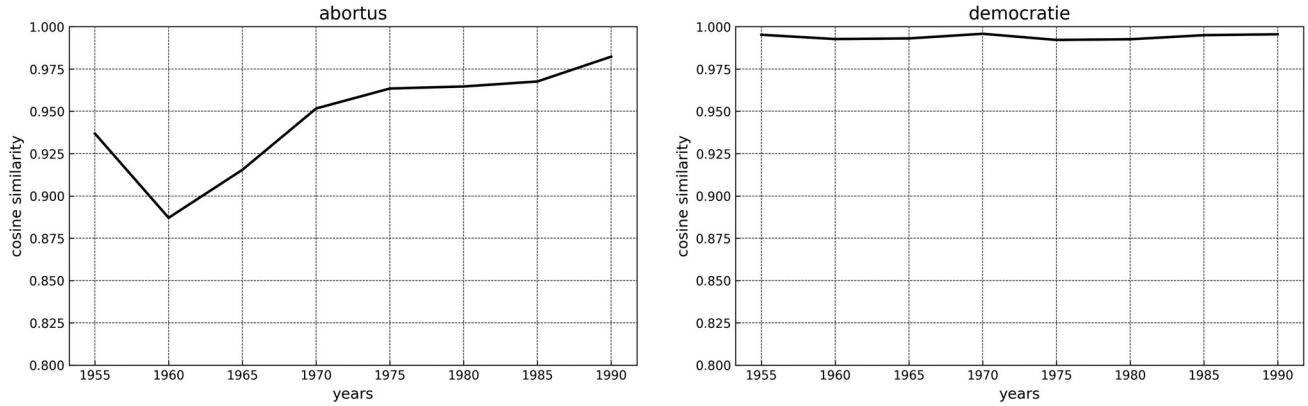

Figure 6. Changes in local neighbors ($k = 25$) of two target words in embeddings trained on historical Dutch Newspapers.

1987 is used in similar contexts as 'Bill Clinton' is used in documents from 1997? Such diachronic word embeddings allow comparative search for entities or topics that are discussed in the same way as a given entity or topic, but in a different period.

Incorporating WEMs in the process of searching and selecting digitized source materials also offers historians a method for dealing with OCR errors (Egense 2017). If the corpus is large enough or the OCR error rate low enough, correctly recognized words would occur in similar contexts as their incorrectly recognized variants. In the WEM, they will be located close to each other, by including close neighbors with small word variations, users can expand query keywords with their misrecognized variants. However, if the error rate is too high, there will be many variant spellings of the same word, each with a low frequency, making it unlikely that they appear in similar contexts. As a consequence, the embedding space will not put misrecognized variants close to the correct word form nor close to other flawed variants.

Query expansion can be seen as a form of concept-based search (Qiu and Frei 1993), where instead of retrieving documents based on query keywords that occur in documents, both documents and queries are mapped via groups of related keywords to something that more directly resembles concepts (Egozi, Markovitch, and Gabrilovich 2011). In this sense, word embeddings also provide researchers with the ability to search for concepts and their related words over time via semantic fields. Rather than tracing individual words, they can trace networks of words, that are constructed by selecting related terms within an embedding space (Kenter et al. 2015).

Martinez-Ortiz et al. present the tool ShiCo (Shifting Concepts), which allows researchers to gain insight into the evolution of concepts and their associations over time (Martinez-Ortiz et al. 2016).[13] For example, when examining the concept of "efficiency" in Dutch public discourse, we are dealing with a relatively recent concept. However, using ShiCo, we can query "efficiency" and using its network of related words, we can trace the concept back in time. This approach allows us to examine the concept "efficiency" even before the word appeared in Dutch public discourse, yielding associated words for particular periods, such as "rationalization', "market forces," and "company returns."

Using WEMs for searching historical corpora has potential, but all the requirements mentioned in Section, "**Considerations before training a model**" and the pitfalls identified in Section, "**Evaluating the quality of a trained word embedding model**" apply. Large amounts of data are needed to train reliable models. In the case of digitized texts, the quality of OCR or HTR needs to be high enough to capture the statistical structure of co-occurring words and their variants.

Conclusion

In this article, we have discussed how historians can assess whether word embeddings might be useful for their research question. We have described three kinds of questions for which WEMs can be useful. First, WEMs can be useful to improve heuristics. The models can be used to expand queries with historically-situated related terms. Second, WEMs can be used for questions related to high-level semantic change. If data size permits, historians can trace semantic change over time and for particular types of documents. This can help to pinpoint change points, but also the semantic fields in which certain words appeared or disappeared. By relying on clear hypotheses, we can avoid cherry-picking results extracted from the models and reading too much into the results. Ideally, results

need to be cross-examined using control words, or by exploring how certain words relate to other words through algebraic operations performed on the vectors. Determining a semantic field is no clear-cut exercise, and it requires domain-knowledge.

Moreover, we have offered several recommendations for training and evaluating models. The size, quality, and bias of the data set have an impact on the reliability of the models. We have also shown that existing evaluation metrics provide some insight into the reliability of the models, but these methods are often not suited for historical data sets. Therefore, they need to be expanded with qualitative assessments and historically-situated lexicons of control words and analogy sets.

One of the most manifest impacts on the usefulness of WEMs is the size of the data. Tracing words and concepts in historical corpora using WEMs is not reliable if the words of interest are relatively rare. The central point of extracting semantics from the statistical structure of text is that there are many occurrences of a word, with a significant overlap in the contexts in which those words appear. Although words with a low frequency are potentially interesting as they are less common and, therefore, less likely to have been studied before, there are only a few contexts in which they appear. For research questions and topics that focus on uncommon words, it is worth considering using different techniques and, more importantly, whether this can be done reliably at all.

This problem is especially stringent for small data sets, where relatively frequent words of interest do not have enough occurrences and contexts to establish stable and reliable embeddings. The same holds for digitized corpora with low-quality OCR/HTR, as the recognition errors distribute the contexts of a word over a large number of variant word forms, leaving no structure to represent meaningful connections.

Recent innovations in Natural Language Processing have produced Deep Contextualized Word Embeddings (Peters et al. 2018). These models enable researchers to not only study semantic changes on a high level, but also on a more fine-grained level that is more sensitive to word ambiguity and more able to capture relatively infrequent word expressions. However, training these models is extremely computationally expensive, and using them is more complicated than using, for example, Word2Vec. Moreover, these models have primarily been trained and used on modern language (Devlin et al. 2019). In future work, we aim to explore how these methods can aid historical inquiry.

Notes

1. https://github.com/mmihaltz/word2vec-GoogleNews-vectors
2. https://www.english-corpora.org/coha/
3. http://www.laurenceanthony.net/software.html
4. The language processing toolkit Gensim offers a fast and easy-to-use implementation of Word2Vec, which further pushed its widespread adoption. https://radimrehurek.com/gensim/index.html
5. For a more in-depth overview of neural networks for Natural Language Processing, see (Goldberg 2015)
6. A concept is described using a single word, but it is not expressed just by this single word but by a set of words and their linkages.
7. For more on this see: (Orlikowski, Hartung, and Cimiano 2018)
8. A popular analogy set can be found here: https://github.com/nicholas-leonard/word2vec/blob/master/questions-words.txt
9. The models used for these experiments are described here: (Wevers 2019)
10. In this paper, the authors do not align the embedding models, and they compare embeddings within each period.
11. https://www.meertens.knaw.nl/nvb/
12. https://historyofwork.iisg.nl/
13. https://github.com/NLeSC/ShiCo

References

Allan, K. and Robinson, J.A. (Eds.). 2012. *Current methods in historical semantics*. Berlin: De Gruyter Mouton.

Andersen, N. Å. 2003. *Discursive analytical strategies: Understanding Foucault, Koselleck, Laclau, Luhmann*. 1st ed., Bristol: Bristol University Press. 10.2307/j.ctt1t898nd.

Antoniak, M., and D. Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics* 6:107–19. doi: 10.1162/tacl_a_00008.

Azarbonyad, H., M. Dehghani, K. Beelen, A. Arkut, M. Marx, and J. Kamps. 2017. Words are malleable: Computing semantic shifts in political and media discourse. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, ACM, 1509–1518. doi: 10.1145/3132847.3132878.

Baroni, M., G. Dinu, and G. Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *ACL* 1 (1):238–247.

Barranco, R. C., R. F. Dos Santos, M. S. Hossain, and M. Akbar. 2018. Tracking the Evolution of Words with Time-reflective Text Representations. 2018 IEEE International Conference on Big Data (Big Data), IEEE, 2088–2097. doi: 10.1109/BigData.2018.8621902.

Batchkarov, M., T. Kober, J. Reffin, J. Weeds, and D. Weir. 2016. A critique of word similarity as a method for evaluating distributional semantic models. Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP, Association for Computational Linguistics, Berlin, Germany, 7–12.

Bingham, A. 2010. The digitization of newspaper archives: Opportunities and challenges for historians. *Twentieth Century British History* 21 (2):225–31. doi: 10.1093/tcbh/hwq007.

Bollmann, M., and A. Søgaard. 2016. Improving historical spelling normalization with bi-directional LSTMs and multi-task learning. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, presented at the COLING 2016, The COLING 2016 Organizing Committee, Osaka, Japan, 131–139.

Bolukbasi, T., K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in neural information processing systems 29*, ed. D.D. Lee, M. Sugiyama, U.V. Luxburg, I. Guyon and R. Garnett, 4349–57. Curran Associates, Inc.

Brunner, O., W. Conze, and R. Koselleck. 2004. *Geschichtliche grundbegriffe: Historisches lexikon zur politisch-sozialen sprache in deutschland. 8 bände in 9*. Stuttgart, Germany: Klett-Cotta.

Carpineto, C., and G. Romano. 2012. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys* 44 (1):1–50. doi: 10.1145/2071389.2071390.

De Bolla, P. 2013. *The architecture of concepts: The historical formation of human rights*. New York: Fordham University Press.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41 (6):391–407. doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), presented at the NAACL-HLT 2019, Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.

Diaz, F., B. Mitra, and N. Craswell. 2016. Query expansion with locally-trained word embeddings. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1, 367–377. doi: 10.18653/v1/P16-1035.

Dubossarsky, H., D. Weinshall, and E. Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 1136–1145.

Efthimiadis, E. N. 1996. Query expansion. *Annual Review of Information Science and Technology (ARIST)* 31:121–87.

Egense, T. 2017. Automated improvement of search in low quality OCR using Word2Vec. https://sbdevel.wordpress.com/2017/02/02/automated-improvement-of-search-in-low-quality-ocr-using-word2vec/.

Egozi, O., S. Markovitch, and E. Gabrilovich. 2011. Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems* 29 (2):1–34. doi: 10.1145/1961209.1961211.

Foucault, M. 2012. *The archaeology of knowledge [1974]*. London: Random House.

Garg, N., L. Schiebinger, D. Jurafsky, and J. Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115 (16):E3635–E3644. doi: 10.1073/pnas.1720347115.

Gavin, M. 2015. "The arithmetic of concepts: A response to Peter de Bolla". *Modeling Literary History*, 18 September. Accessed February 25, 2019 http://modelingliteraryhistory.org/2015/09/18/the-arithmetic-of-concepts-a-response-to-peter-de-bolla/.

Giuliano, V. E., and P. E. Jones. 1962. "Linear associative information retrieval". *Computer Science*. 10.21236/ad0290313.

Hamilton, W. L., J. Leskovec, and D. Jurafsky. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, Vol. 2016, NIH Public Access, 2116.

Hamilton, W. L., J. Leskovec, and D. Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), presented at the Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 1489–1501. doi: 10.18653/v1/P16-1141.

Hamilton, W. L., J. Leskovec, and D. Jurafsky. 2016c. Cultural shift or linguistic drift? comparing two computational measures of semantic change. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, Vol. 2016, NIH Public Access, 2116.

Harris, Z. S. 1954. Distributional structure. *Word* 10 (2-3): 146–62. Nodoi: 10.1080/00437956.1954.11659520.

Hellrich, J., S. Buechel, and U. Hahn. 2018. JeSemE: Interleaving Semantics and Emotions in a Web Service for the Exploration of Language Change Phenomena. Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Santa Fe, New Mexico, 10–14.

Hellrich, J., and U. Hahn. 2016. Bad Company-Neighborhoods in Neural Embedding Spaces Considered Harmful. *COLING* 2785–96.

Huistra, H., and B. Mellink. 2016. Phrasing history: Selecting sources in digital repositories. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 49 (4):220–9. doi: 10.1080/01615440.2016.1205964.

Ifversen, J. 2003. Text, discourse, concept: Approaches to textual analysis. *Kontur* 7:60–9.

James, P., and M. B. Steger. 2014. A Genealogy of 'Globalization': The Career of a Concept. *Globalizations* 11 (4):417–34. doi: 10.1080/14747731.2014.951186.

Junge, K., and K. Postoutenko. 2014. *Asymmetrical concepts after reinhart koselleck: Historical semantics and beyond*. Bielefeld: Transcript Verlag.

Jurafsky, D., and J. Martin. 2009. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River: Pearson.

Kenter, T., M. Wevers, P. Huijnen, and M. de Rijke. 2015. Ad Hoc monitoring of vocabulary shifts over time. Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, ACM, New York, 1191–1200. doi: 10.1145/2806416.2806474.

Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D. and Petrov, S. (2014), "Temporal analysis of language through Neural Language Models", *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, Association for Computational Linguistics, Baltimore, MD, USA, pp. 61–65.

Koolen, M., F. Adriaans, J. Kamps, and M. De Rijke. 2006. A cross-language approach to historic document retrieval. *European Conference on Information Retrieval*, 407–19.

Koselleck, R. 1989. Social History and Conceptual History. *International Journal of Politics, Culture and Society* 2 (3): 308–25. doi: 10.1007/BF01384827.

Koselleck, R. 2002. *The practice of conceptual history: Timing history, spacing concepts*. 1st ed. Stanford: Stanford University Press.

van Lange, M., and R. Futselaar. 2018. Debating evil: Using word embeddings to analyze parliamentary debates on war criminals in The Netherlands. Proceedings of the Conference on Language Technologies & Digital Humanities, 147–153.

Levy, O., and Y. Goldberg. 2014. Linguistic Regularities in Sparse and Explicit Word Representations. Proceedings of the Eighteenth Conference on Computational Natural Language Learning, presented at the Proceedings of the Eighteenth Conference on Computational Natural Language Learning, Association for Computational Linguistics, Ann Arbor, Michigan, 171–180. doi: 10.3115/v1/W14-1618.

Levy, O., Y. Goldberg, and I. Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics* 3:211–25. doi: 10.1162/tacl_a_00134.

Martinez-Ortiz, C., T. Kenter, M. Wevers, P. Huijnen, J. Verheul, and J. Van Eijnatten. 2016. Design and implementation of ShiCo: Visualising shifting concepts over time. *HistoInformatics 2016* 1632:11–9.

Matsuoka, J., and Y. Lepage. 2014. Measuring similarity from word pair matrices with syntagmatic and paradigmatic associations. Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex), Association for Computational Linguistics and Dublin City University, Dublin, Ireland, 77–86. doi: 10.3115/v1/W14-4712.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in neural information processing systems 26*, ed. C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K.Q. Weinberger, 3111–9. Curran Associates, Inc.

Mikolov, T., W. Yih, and G. Zweig. 2013. Linguistic regularities in continuous space word representations. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 746–51.

Miller, G. A., and W. G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6 (1):1–28. doi: 10.1080/01690969108406936.

Nicholson, B. 2013. The digital turn: Exploring the methodological possibilities of digital newspaper archives. *Media History* 19 (1):59–73. doi: 10.1080/13688804.2012.752963.

Orlikowski, M., M. Hartung, and P. Cimiano. 2018. Learning diachronic analogies to analyze concept change. Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, 1–11.

Palti, E. J. Reinhart Koselleck: His concept of the concept and Neo-Kantianism. *Contributions to the History of Concepts* 6 (2):1–20. doi:10.3167/choc.2011.060201.

Palti, E. J. 2011. Reinhart Koselleck: His Concept of the Concept and Neo-Kantianism, *Contributions to the History of Concepts*, 6 (2): 1–20.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L. 2018, "Deep Contextualized Word Representations", *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, presented at the NAACL-HLT 2018, Association for Computational Linguistics, New Orleans, Louisiana, pp. 2227–2237.

Piotrowski, M. 2012. Natural language processing for historical texts. *Synthesis Lectures on Human Language Technologies* 5 (2):1–157. doi: 10.2200/S00436ED1V01Y201207HLT017.

Qiu, Y., and H.-P. Frei. 1993. Concept based query expansion. Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 160–169. doi: 10.1145/160688.160713.

Rapp, R. 2003. Syntagmatic and paradigmatic associations in information retrieval. In *Between data science and applied data analysis*, ed. M. Schader, W. Gaul and M. Vichi, 473–82. Berlin: Springer Berlin Heidelberg.

Recchia, G., E. Jones, P. Nulty, J. Regan, and P. de Bolla. 2016. Tracing shifting conceptual vocabularies through time. *European knowledge acquisition workshop*, 19–28. Springer.

Richter, C., M. Wickes, D. Beser, and M. Marcus. 2018. Low-resource post processing of noisy OCR output for historical corpus digitisation. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), presented at the LREC 2018, European Language Resources Association (ELRA), Miyazaki, Japan. Accessed April 13, 2020. https://www.aclweb.org/anthology/L18-1369.

Richter, M. 1987. Begriffsgeschichte and the History of Ideas. *Journal of the History of Ideas* 48 (2):247–63. doi: 10.2307/2709557.

Roy, D., D. Paul, M. Mitra, and U. Garain. 2016. Using word embeddings for automatic query expansion. *ArXiv Preprint ArXiv:1606.07608*

Rudolph, M., and D. Blei. 2018. Dynamic embeddings for language evolution. Proceedings of the 2018 World Wide Web Conference, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1003–1011. doi: 10.1145/3178876.3185999.

Sahlgren, M., and A. Lenci. 2016. The Effects of Data Size and Frequency Range on Distributional Semantic Models. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, presented at the EMNLP 2016, Association for Computational Linguistics, Austin, Texas, 975–980.

Salton, G., A. Wong, and C.-S. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM* 18 (11):613–20. doi: 10.1145/361219.361220.

Schmieder, F. 2019. "Begriffsgeschichte's methodological neighbors and the scientification of concepts", 2 October. Accessed October 28, 2019. https://jhiblog.org/2019/10/02/begriffsgeschichtes-methodological-neighbors-and-the-scientification-of-concepts/.

Schütze, H., and J. Pedersen. 1993. A vector model for syntagmatic and paradigmatic relatedness. *Proceedings of the 9th Annual Conference of the UW Centre for the New OED and Text Research* :104–13.

Skinner, Q. 2002. *Visions of politics*. Cambridge: Cambridge University Press.

Sommerauer, P., and A. Fokkens. 2019. Conceptual change and distributional semantic models: An exploratory study on pitfalls and possibilities. Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, presented at the Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, Association for Computational Linguistics, Florence, Italy, 223–233. doi: 10.18653/v1/W19-4728.

Sommerauer, P. J. M., and A. S. Fokkens. 2018. Firearms and tigers are dangerous, kitchen knives and zebras are not: Testing whether word embeddings can tell. The 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP: Proceedings of the First Workshop, Association for Computational Linguistics (ACL), 276–286.

Szymanski, T. 2017. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 448–453.

Tulkens, S., C. Emmery, W. Daelemans, et al. 2016. Evaluating unsupervised dutch word embeddings as a linguistic resource. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), ed. N.C Chair, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno. European Language Resources Association (ELRA).

Wang, B., A. Wang, F. Chen, Y. Wang, and C.-C. J. Kuo. 2019. Evaluating word embedding models: Methods and experimental results. In *APSIPA transactions on signal and information processing*, vol. 8. Cambridge: Cambridge University Press. doi: 10.1017/ATSIP.2019.12.

Wevers, M. 2019. Using word embeddings to examine gender bias in dutch newspapers, 1950-1990. In Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, presented at the ACL, ed. N. Tahmasebi, L. Borin, A. Jatowt and Y. Xu. Florence, Italy. doi: 10.18653/v1/W19-4712.

Wittgenstein, L. 2010. *Philosophical investigations*. Hoboken, New Jersey: John Wiley & Sons.

Yao, Z., Y. Sun, W. Ding, N. Rao, and H. Xiong. 2018. Dynamic Word Embeddings for Evolving Semantic Discovery. Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining - WSDM '18, 673–681. doi: 10.1145/3159652.3159703.