Chapter 11

Challenges for computational lexical semantic change

Simon Hengchen[a], Nina Tahmasebi[a], Dominik Schlechtweg[b] & Haim Dubossarsky[c]
[a]University of Gothenburg [b]University of Stuttgart [c]University of Cambridge

The computational study of lexical semantic change (LSC) has taken off in the past few years and we are seeing increasing interest in the field, from both computational sciences and linguistics. Most of the research so far has focused on methods for modelling and detecting semantic change using large diachronic textual data, with the majority of the approaches employing neural embeddings. While methods that offer easy modelling of diachronic text are one of the main reasons for the spiking interest in LSC, neural models leave many aspects of the problem unsolved. The field has several open and complex challenges. In this chapter, we aim to describe the most important of these challenges and outline future directions.

1 Introduction

The goal of tackling lexical semantic change (LSC) computationally is primarily to reconstruct semantic change evident in large diachronic corpora. The first papers addressing LSC appeared in 2008–2009 and since then a few papers per year have been published.[1] The first works that used neural embeddings were published in 2014 and 2015 (Kim et al. 2014, Kulkarni et al. 2015, Dubossarsky et al. 2015) and together with Hamilton et al. (2016b), they sparked interest in the

[1] “Language evolution”, “terminology evolution”, “semantic change”, “semantic shift” and ’‘semantic drift” are all terms that are or have been used for the concept which we denote *lexical semantic change.*

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

research community in the problem of LSC.[2] Although there are a few research groups that have a longer history of studying LSC using computational methods, the majority of papers are single-entry papers where a group with an interesting model apply their method to a novel application on popular diachronic data. This leads to quick enhancement of methods but limits development and progress in other aspects of the field.

When surveying prior work, it is obvious that the computational field of LSC has been divided into two strands. The first strand deals with words as a whole and determines change on the basis of a word's dominant sense (e.g. Kim et al. 2014, Kulkarni et al. 2015). An oft-used example is *gay*[3] shifting from its 'cheerful' sense to 'homosexual'. The second strand deals with a word's senses[4] individually – for example, the 'music' sense of *rock* has gradually come to describe not only music but also a certain lifestyle, while the 'stone' sense remained unchanged (as seen in the works of Tahmasebi 2013 and Mitra et al. 2015). The first strand took off with the introduction of neural embeddings and its easy modelling of a word's semantic information. The second strand, faced with the immense complexity of explicitly modelling senses and meaning, has received much less attention.

Computational models of meaning are at the core of LSC research, regardless of which strand is chosen. All current models, with the exception of those purely based on frequency, rely on the distributional hypothesis, which brings with it the set of challenges discussed in Section 3. But even accepting the distributional hypothesis and assuming meaning in context, the problem formulated by Schütze (1998) remains: how does one accurately portray a word's senses? The question is valid regardless of whether the senses are represented individually or bundled up into one single representation. Recent developments in contextual embeddings (e.g. Peters et al. 2018) provide hope for accurate modelling of senses. However, they do not alleviate the problem of grouping sentence representations into sense correspondence. Within natural language processing (NLP), computational models of word meaning are often taken at face value and not questioned

[2] Compare, for example, the roughly 30 papers at the start of 2018 as reported by Tahmasebi et al. (2018), with the roughly 50 papers submitted at the 1st International workshop on wom-putational approaches to historical language change 2019 (Tahmasebi et al. 2019), and recent submissions at xACL venues, including the 21 papers submitted to the SemEval-2020 Task 1 on unsupervised lexical semantic change detection (Schlechtweg et al. 2020).

[3] More often than not, parts of speech are collapsed – in this case, there is thus no difference between the adjective and the noun.

[4] For the sake of clarity we use this as a simplified wording. We do not imply that a fixed number of senses exist in a sense inventory; instead senses can overlap and be assigned different strengths.

by researchers working on LSC. This is thus one of the areas that needs further attention in future work. Another area for thorough investigation is how useful sense-differentiation is for accurate LSC models.

Another important area for future work is robust evaluation. Computational LSC methods model textual data as information signals and detect change in these signals. A signal can be a multidimensional vector, a cluster of words, topics, or frequency counts. This increasing level of abstraction is often ignored in evaluation; current evaluation standards allow for anecdotal evaluation of signal change, often without tying the results back to the text. Can we find evidence in the text for the detected changes?[5] So far, semantic annotation is the only way to evaluate methods on historical corpora while making sure that expected changes are present in the text. Annotating involves a significant investment of time and funds, and results in a limited test set. A middle ground is to evaluate with respect to an outside source, like a dictionary or encyclopedia. However, while these resources offer an “expected” time and type of change, we can never be certain that these changes are reflected in the corpus under study. We refer to the example of *computer* in Tahmasebi et al. (2018): the different parts of Google Books (British English, American English, and German) that reach the same level of frequency in different periods in time, 1934 for the German portion, 1943 for the American English, and 1953 for the British English. Recent work (Kulkarni et al. 2015, Dubossarsky et al. 2019, Shoemark et al. 2019, Schlechtweg & Schulte im Walde 2020) introduced relatively cheap methods of generating *synthetic* semantic change for any dataset, which we believe is an important path forward. The question is not if, but how synthetic evaluation data can complement costly manual evaluation. This will be answered in Section 4.

In addition, current methods work with rather coarse time granularity, largely because of the inherent complexity of adding multiple time bins (and senses) to the models. Unfortunately this constraint limits both the possibilities of the methods, and the results that can be found. Again, adding complexity to the models results in complexity in evaluation and calls for robust evaluation methods and data.

In this chapter, we will discuss current and future challenges, and outline avenues for future work. More specifically, Section 2 discusses the requirements for textual resources, and their role for LSC. Section 3 covers models of meaning,

[5] To the best of our knowledge, only Hengchen (2017) evaluates semantic change candidates output by a system reading a relatively large sample of sentences from the corpus studied – but only for a single word, while several projects make use of extensive annotation to ensure that detected changes are present in the underlying textual corpus (Lau et al. 2012, Schlechtweg et al. 2017, 2018, 2020, Hätty et al. 2019, Perrone et al. 2019, Giulianelli et al. 2020).

the current limitations of computational models, models of change, and what remains to be done in terms of time and sense complexity. Section 4 sheds light on the need for robust evaluation practices and what we have learned so far. We then proceed in Section 5 to showing the potential for collaboration between computational LSC and other fields, and end with some concluding remarks.

2 Data for detecting LSC

Hand in hand with the fast and simple modelling of word meaning, using neural embeddings for example, is the easy access to digital, diachronic texts that sparked mainstream interest in LSC as a problem domain for testing new models. For many reasons, including the early availability of large English corpora, there has long been a large over-representation of studies performed on English, in particular using COHA (Davies 2002), Google N-grams (Michel et al. 2011), and various Twitter corpora (see Table 2 in Tahmasebi et al. 2018 for an overview). As a consequence, most of the computational modelling of LSC has been developed and evaluated on these resources. However, tools and methods developed on one language (e.g., English) are not easily transferable to another language, a reoccurring challenge in other fields of NLP as well (Bender 2011, Ponti et al. 2019). Moreover, many languages may even lack the amount of historical, digitized data needed for robustly employing state-of-the-art methods like neural embeddings. This point is reinforced if we follow the recommendations of Bowern (2019) for example, and distinguish groups based on features such as location, age, and social standing. The immediate result of this limitation is that many languages remain unstudied, or worse, studied with unsuitable methods. Consequently, whatever conclusions are drawn about LSC and formulated as “laws” are based on a very limited sample of languages, which may not be representative of the other 7,100 living languages.[6] As LSC mainly focuses on diachronic text, one obvious area for research is determining how well methods that rely on representations developed primarily for modern text transfer to historical languages.[7]

[6] Figure from ethnologue.com, https://www.ethnologue.com/guides/how-many-languages, last accessed 2020/01/16, rounded down to the lower hundred. In addition to living languages, dead languages are also studied, e.g. Rodda et al. (2017), Perrone et al. (2019), or McGillivray et al. (2019) for Ancient Greek.

[7] See Piotrowski (2012) for a thorough overview of such challenges, and Tahmasebi et al. (2013) for an example of the applicability of a cluster-based representation on historical data.

An important criterion for textual resources for LSC is good, reliable time stamps for each text. The texts should also be well distributed over longer time periods.[8] For these reasons, newspaper corpora are popular for LSC studies. They also have the advantage that different news items are roughly equal in length. But while they cover a large and interesting part of society, including technological inventions, they are limited in their coverage of everyday, non-newsworthy life.[9] On the other hand, large literary corpora like the Google N-grams pose different challenges including their skewed sampling of topics over time. For example, Pechenick et al. (2015) point to an over-representation of scientific literature in the Google Books corpus, which biases the language used toward specific features mostly present in academic writing. A second challenge to the use of the Google N-grams corpus is the small contexts (at most five words in a row) and the scrambled order in which these contexts are presented. To what extent this large resource can be used to study LSC remains to be investigated. One important question is whether changes found can be thoroughly evaluated using only these limited contexts.[10]

Other, less known corpora have other known deficits. For example, many literary works come in multiple editions with (minor) updates that modernize the language, while other texts lack known timestamps or publication dates. Some authors are more popular than others (or were simply more productive) and thus contribute in larger proportion and risk skewing the results.[11]

What an optimal resource looks like is clearly dependent on the goal of the research, and LSC research is not homogeneous; different projects have different aims. While some aim to describe an interesting dataset (like the progression of one author), others want to use large-scale data to generalize to language outside of the corpus itself. In the latter case, it is important to be varied, as large textual

[8] Recurring advertisements running for weeks at a time can effectively bias the corpus. See for example, the work by Prescott (2018: 67) for a case study on the Burney newspaper collection.

[9] Similar to news corpora, Twitter, another popular source for LSC research, offers posts which have timestamps and are consistent in size (though radically shorter than news articles), but with very different characteristics. However, most Twitter corpora are short-term, unlike the longer temporal dimensions of many news corpora.

[10] Although there are several LSC papers using Google N-grams, e.g., Wijaya & Yeniterzi (2011) or Gulordava & Baroni (2011), to date there are no systematic investigations into the possibility of detecting different kinds of LSC, nor any systematic evaluation using grounding of found change in the Google N-grams.

[11] See, for example, Tangherlini & Leonard (2013) where a single book completely changed the interpretation of a topic. Similarly, one can find many editions and reprints of the Bible in the Eighteenth Century Collections Online (ECCO) dataset that spans a century and contains over 180,000 titles; which will influence models. For a study on the effects of text duplication on semantic models, see Schofield, Thompson, et al. (2017).

corpora are not random samples of language as a whole (see, e.g., Koplenig 2016) and whatever is encountered in corpora is only valid for those corpora and not for language in general.

Most existing diachronic corpora typically grow in volume over time. Sometimes this stems from the amount of data available. At other times it is an artefact related to ease of digitisation (e.g., only certain books can be fed into an automatic scanner) and OCR technology (OCR engines are trained on specific font families). This growth results in an extension of vocabulary size over time, which might not reflect reality and has serious effects on our methods. For example, previous work has shown that diachronic embeddings are very noisy (Hellrich & Hahn 2016, Dubossarsky et al. 2017, 2019, Kaiser et al. 2020, Schlechtweg et al. 2020) with a large frequency bias and are clearly affected by more and more data over time. Current and future studies are thus left with the question of whether the signal change we find really correspond to LSC in the text, or whether it is simply an artefact of the corpus.

We can only find what is available in our data: if we want to model other aspects of language and LSC, we need datasets that reflect those aspects well (for similar considerations related to evaluation, see Section 4.1.1). This fact makes the case for using texts stemming from different sources, times, and places to allow for (re-)creating the complex pictures of semantic change. Thus general aspects beneficial for LSC are texts that are well-balanced (in time and across sources) and high-quality (with respect to OCR quality) with clear and fine-grained temporal metadata, as well as other kinds of metadata that can be of use. Until now, most existing computational LSC studies have been performed on textual data exclusively, aside from Perrone et al. (2019) and Jawahar & Seddah (2019) who respectively used literary genre and social features as features. The reason for the under-utilisation of extra-linguistic metadata – despite there being great need for it, as advocated by Bowern (2019) – is to a large extent the lack of proper and reliable metadata. In the case of Google N-grams, this kind of metadata is sacrificed in favour of releasing large volumes of data freely. This path is also promising with respect to modelling the individual intent, described in Section 3.3.1.

For the long-term future, we should raise the question of whether we can model language at all using only texts (Bender & Koller 2020). How much can we improve with multi-modal data in the future (Bruni et al. 2012), and what kind of data would be beneficial for LSC?

3 Models of meaning and meaning change

In this section, we shed light on the "meaning" we strive to model from the data, and on the challenges involved with modelling meaning computationally, and finally on how to employ the resulting information signals to establish change.

3.1 Theory of lexical meaning and meaning change

The field urgently needs definitions of the basic concepts it wants to distinguish: after all, we can draw from a rich tradition of semantics and semantic change research. The field traditionally starts with Reisig (1839), although Aristotle (1898) theorized metaphors in his *Poetics* well before then.[12]

Here we focus on one theory which encompasses many others. Blank (1997: 54) distinguishes three different levels of word meaning based on which type of knowledge a word can trigger in a human: (i) language-specific semantic, (ii) language-specific lexical, and (iii) language-external knowledge. The first comprises core semantic knowledge needed to distinguish different word meanings from each other.[13] This knowledge corresponds to the minimal language-specific semantic attributes needed to structure a particular language, often called "sememe" in structural semantics. From these follow the hierarchical lexical relations between words (e.g. synonymy or hypernymy). The second level of word meaning comprises knowledge about the word's role in the lexicon (part of speech, word family or knowledge about polysemy/multiple meanings, referred to as "senses" in this chapter). It includes the rules of its use (regional, social, stylistic or diachronic variety; syntagmatic knowledge such as selectional restrictions, phraseologisms or collocations). Level (iii) comprises knowledge about connotation and general knowledge of the world.

Blank (1997) assumes that the knowledge from these three levels is stored in the mental lexicon of speakers, which can also change historically in these three levels (at least). An example of a change at the language-specific semantic level (i) is Latin *pipio* 'young bird' > 'young pigeon' which gained the attribute [pigeon-like] (Blank 1997: 106–107). An example of change on the language-specific lexical level (ii) is *gota* 'cheek' which changes from being commonly used in Old Italian to being used exclusively in the literary-poetic register in New Italian (Blank 1997: 107). Finally, a change at the language-external knowledge level (iii) occurs when the knowledge about the referent changes. This can occur, for example, when the

[12] See for example the work by Magué (2005) for an overview.

[13] Note that this level covers only what Blank (1997) calls "knowledge" (p. 94). He then distinguishes six further levels of "meaning" (pp. 94–96).

referent itself changes such as with German *Schiff* 'ship', as ships were primarily steamships in the 19th century, while today they are mainly motor ships (Blank 1997: 111).

Unfortunately, as will be made clear in the following subsection, this rich tradition of work is not used by the computational side of LSC because it is difficult to model meaning purely from written text. Currently, our modelling is very blunt. It can primarily capture contextual similarity between lexical items, and rarely distinguishes between different levels of meaning. Whether we draw from the large existing body of work that exists in traditional semantics research, or start from scratch with a new definition of what computational meaning is, we hope researchers in our field can come together and agree on what is, and should, be modelled.

Similar to the conundrum in the definition of word meaning above, studies on LSC detection are seldom clear on the question of which type of information they aim to detect change in (Schlechtweg & Schulte im Walde 2020). There are various possible applications of LSC detection methods (e.g. Hamilton et al. 2016a, Voigt et al. 2017, Kutuzov et al. 2017, Hengchen et al. 2019). Change at different levels of meaning may be important for different applications. For example, for literary studies it may be more relevant to detect changes of style, for social sciences the relevant level may be language-external knowledge and for historical linguistics the language-specific lexical and semantic levels may be more important. Furthermore, LSC can be further divided into types (e.g., broadening/narrowing, amelioration/pejoration, metaphor and hyperbole). Several taxonomies of change have been suggested over the decades (Bréal 1897, Bloomfield 1933 and Blank 1999, to name a few). Clearly, none of these applications or types of change can be properly tested until an adequate model of meaning is developed and the types of LSC to be investigated are meticulously defined.

3.2 Computational models of meaning

The need to choose the textual data available for the models and the decisions regarding the preprocessing of the text are common to all models of computational meaning. While the influence of the former was described in Section 2, and is fairly straightforward, extremely little attention is paid to preprocessing although its effects on the end results are far-reaching. The lower-casing of words often conflates parts of speech. For example *Apple* (proper noun) and *apple* (common noun) cannot be distinguished after lower-casing. Filtering out different parts of speech is also common practice, and can have radical effects on the results.[14] Thus the effects of preprocessing on meaning representations should be investigated in the future.

Nevertheless, the core of studying LSC computationally is the choice of the computational model of meaning: what we can model determines what change we can find. Crucially, methods for computational meaning inherit the theoretical limitations discussed in Section 3.1. The challenge becomes even more cumbersome as existing methods for computational meaning rely on the distributional hypothesis (Harris 1954), which represents word meaning based on the context in which words appear. In so doing, they often conflate lexical meaning with cultural and topical information available in the corpus used as a basis for the model. These limitations are not specific to semantic change, and lie at the basis of a heated debate that questions the fundamental capacity of computational models to capture meaning using only textual data, see for example Bender & Koller (2020).

There are different categories of computational models for meaning. These comprise a hierarchy with respect to the granularity of their sense/topic/concept representations:

(a) a single representation for a word and all its semantic information (e.g., static embeddings),

(b) a representation that splits a word into semantic areas (roughly) approximating senses (e.g., topic models), and

(c) a representation that models every occurrence of a word individually (e.g., contextual embeddings) and possibly groups them post-hoc into clusters of semantically-related uses expressing the same sense.

These categories of models differ with respect to their potential to address various LSC problems. For example, novel senses are hard to detect with models of category (a). However, these models have the upside of producing representations for all the words in the vocabulary, which is not the case for all models in category (b) (see for example Tahmasebi et al. 2013). In contrast, some sense-differentiated methods (category (b)), such as topic modelling allow for easy disambiguation so that we can deduce which word was used in which sense. However, category (a) models (e.g., word2vec) do not offer the same capability as

[14] For discussions on the effects of preprocessing (or, as coined by Thompson & Mimno 2018, "purposeful data modification") for text mining purposes, we refer to Schofield & Mimno (2016), Schofield, Magnusson, et al. (2017), Denny & Spirling (2018), and Tahmasebi & Hengchen (2019).

they provide one vector per word, which is also biased toward the word's more frequent sense.[15]

Furthermore, models that derive meaning representations that can be interpreted and understood are needed to determine which senses of a word are represented, and whether they capture standard word meaning, topical use, pragmatics, or connotation (i.e., to distinguish between the levels of meaning referred to in Section 3.1). The interpretability also allows us to qualitatively investigate different representations to determine which is better for different goals.

Finally, the data requirements of our models can pose a critical limitation on our ability to model meaning (see Section 2) as the computational models of meaning are data hungry and require extremely large amounts of text. They cannot be applied to the majority of the world's existing written languages as those often do not have sufficient amounts of written historical texts. If we follow the proposal of Bowern (2019) and divide our data, not only by time, but also according to social aspects (to e.g. echo Meillet 1905), we reduce the amount of available data even further.

3.3 Computational models of meaning change

Change is defined and computed by comparing word representations between two or more time points, regardless of the specific model of meaning. Different models entail different mathematical functions to quantify the change. For example, and without claiming to be exhaustive: cosine or Euclidean distances are used for embedding models of continuous vectors representations, Hellinger distance and Jensen-Shannon or Kullback-Leibler divergences for topic distributions, and Jensen-Shannon divergence or cross-entropy for sense-differentiated representations. Ultimately the mathematical functions provide only a scalar that represents the degree of change. Determining change type from this information is not straightforward. This impedes our ability to derive fine-grained information about the nature of change, as touched upon in Section 3.1, and to incorporate theories of change which, for instance, postulate direction of change. For example, it becomes difficult to detect which sense changed, or to provide relevant distinctions related to the different applications (e.g., change in meaning vs. change in connotation), or taxonomies of change (e.g., broadening vs. narrowing). Schlechtweg & Schulte im Walde (2020) identified two basic notions of change that are

[15] Every use of the word in a sentence is not accurately described by the vector representing it. In modern texts not pertaining to geology, a vector representation of *rock* is biased toward its more frequent 'music' sense and will be a worse representation of a sentence where *rock* is used in a 'stone' sense.

used in LSC research to evaluate the models' scalar change scores: (i) GRADED LSC, where systems output to what degree words change (Hamilton et al. 2016a, Dubossarsky et al. 2017, Bamler & Mandt 2017, Rosenfeld & Erk 2018, Rudolph & Blei 2018, Schlechtweg et al. 2018), and (ii) BINARY LSC, where systems make a decision on whether words have changed or not (Cook et al. 2014, Tahmasebi & Risse 2017, Perrone et al. 2019, Shoemark et al. 2019). Despite this limitation of coarse scores of LSC, several change types have been targeted in previous research (Tahmasebi et al. 2018: Table 4). Importantly, in order for the results of LSC methods to be valuable for downstream tasks, we see a great need to determine the kind of change (e.g., broadening, narrowing, or novel sense). Methods that only detect one class, namely *changed*, defer the problem to follow-up tasks: in which way has a word changed, or on what level (i)–(iii)[16] from Section 3.1 did the change occur?

3.3.1 Discriminating individual sentences

Meaning is ascribed to words at the sentence (utterance) level. However, for technical reasons related to the limitations of current computational models, previous work has carried out LSC only in large corpora. As a result, we model each word of interest with a signal (topic, cluster, vector) across all sentences and detect change in the signal. This discrepancy between the level at which the linguistic phenomena occur and the level of the analysis that is carried out may account for the type of questions commonly asked in contemporary research. In the majority of the cases, the signal change is evaluated on its own, and the question *did the word meaning change or not?* is the only one answered. In a few rare cases, change is tied to the text and verified using the text. *Did the word change in the underlying corpus or not?* is in fact a much more accurate question but is asked much less frequently. In a future scenario, where our models of computational meaning are much more fine-grained, we will be able to ask a third question: *Is a specific usage of a word different than its previous uses?* To be able to tie the detected changes back to individual usage is much more demanding of any system and requires word sense disambiguation (WSD) to be fully solved. Although radically more challenging, this task is also much more rewarding. It can help us in proper search scenarios, in dialogue and interaction studies, argument mining (where a person's understanding of a concept changes during the conversation), and in literary studies, to name but a few examples.

[16] Though level (iii) relates to change in world-knowledge and goes well beyond semantic change.

3.3.2 Modelling of time

The modelling of meaning change is directly dependent on the time dimension inherent in the data. Often, we artificially pool texts from adjacent years into long time bins because our computational models require large samples of text to produce accurate meaning representations or, to draw from research in historical sociolinguistics, because bins of a certain length are considered as "generations" of language users (Säily 2016). Unfortunately, this leads to loss of fine-grained temporal information. From a modelling perspective, the inclusion of such information has the clear advantage of leading to more ecological models for LSC. This advantage can be used in two main ways: either to mitigate the noise associated with meaning representation models, or to detect regular patterns of change. Understandably, these advantages are only available when sufficient time points are included in the analysis. More time points, however, undoubtedly lead to greater computational complexity – linearly if we consider the comparison of only subsequent time points, or quadratically if we consider all pairwise comparisons.

Some theories of LSC assume that change unfolds gradually through time, creating a trajectory of change (e.g., the regular patterns of semantic change in Traugott & Dasher 2001). Only models that acquire a meaning representation at several time points (e.g. Tsakalidis & Liakata 2020) are able to validate this underlying assumption by demonstrating a gradual trajectory of change. The work by Rosenfeld & Erk (2018) is an interesting example, as it models semantic change as a continuous variable and can also output the rate of change. Good extensions include allowing different change rates for different categories of words, or including background information about time periods where things change differently. In addition, models with multiple time points may contribute to improved LSC modelling by facilitating the discovery of intricate change patterns that would otherwise go unnoticed. For example, Shoemark et al. (2019) analysed Twitter data with high temporal resolution, and reported that several words demonstrated repeating seasonal patterns of change. The analysis of LSC trajectories easily lends itself to the use of modern change detection methods, which holds great promise for detecting hidden patterns of both change and regularities.

4 Evaluation

Thus far, evaluation of LSC methods has predominantly ranged from a few anecdotally discussed examples to semi-large evaluation on (synthetic or pre-compiled) test sets, as made clear by Table 2 in Tahmasebi et al. (2018).[17] The SemEval-2020 Task 1 on unsupervised lexical semantic change detection provided the first larger-scale, openly available dataset with high-quality, hand-labeled judgements. It facilitated the first comparison of systems on established corpora, tasks, and gold-labels (Schlechtweg et al. 2020).

However, despite being the largest and broadest existing evaluation framework, the definition of LSC used in Schlechtweg et al. (2020) – i.e., a binary classification and a ranking task – is a radical reduction of the full LSC task. The definition of LSC involves modelling of words and detecting (sense) changes, as well as generalising across many more time points, and disambiguating instances of words in the text. There cannot be only one universal model of a word: there are many ways to describe a word and its senses (see, for example, different dictionary definitions of the same word). So how do we devise evaluation data and methods such that different ways of defining meaning are taken into consideration when evaluating? Should a future evaluation dataset involve dividing the original sentences where a word is used in a particular sense into clusters with sentences that contributed to each sense, to avoid having to evaluate the different representations modelled for a word? How do we handle the uncertainty of which sense led to another? And how many new instances of change are needed to constitute semantic change?

Current work in unsupervised LSC is primarily limited to binary decisions of "change" or "no change" for each word. However, some go beyond the binary to include graded change (although these changes are then often used in figures, for illustrative purposes), and a possible classification of change type. Future work in LSC needs to include a discussion of what role the modelling of sense and signal should play in the evaluation of semantic change: how large does the correspondence between the model and the "truth" for a model need to be, for the results to be deemed accurate? Should we be satisfied to see our methods performing well on follow-up (or downstream) tasks but failing to give proper semantic representation? Evaluation heavily depends on task definition – and thus on the principle of fitness for use.[18] In addition, to study LSC with different task definitions we need to have datasets that reflect these perspectives and make use of task-specific definitions of both meaning and change during evaluation.

[17] The work by Hu et al. (2019) uses *dated* entries of the *Oxford English Dictionary* and thus provides an exception.

[18] A concept originally from Joseph M. Juran, and thoroughly discussed in Boydens (1999).

4.1 Types of evaluations

In the following subsections, we tackle two types of evaluation typically employed for LSC. We first discuss evaluation on ground-truth data, then tackle the promising evaluation on artificially-induced LSC data and argue that both should be used in a complementary fashion.

4.1.1 Ground truth

An important part of evaluation is determining what to evaluate. For example, some studies perform quantitative evaluation of regularities in the vocabulary as a whole. Regardless of other potential evaluation strategies, all existing work (also) evaluates change detected for a small number of lexical items – typically words – in a qualitative manner. This is done in one of two ways: either (i) a set of predetermined words are used for which there is an expected pattern of change, or (ii) the (ranked) output of the investigated method or methods is evaluated. Both of these evaluation strategies have the same aim, but with different (dis)advantages, which we discuss below.

(i) This evaluation strategy consists of creating a pre-chosen test set and has the advantage of requiring less effort as it removes the need to conduct a new evaluation for each change made to parameters such as size of time bins, or preprocessing procedure. The downside is, however, that the evaluation does not allow for new, previously unseen examples.[19] The pre-chosen words can be positive examples (words known to have changed), or negative examples (words known to be stable). Evaluation on only one class of words, positive or negative, does not properly measure the performance of a method. Let us say that we have a method that always predicts change, and we only evaluate on words that have changed. Unless we also evaluate exactly how the word has changed, or when, the method will always be 100% accurate. The best indicator of a method's performance is its ability to separate between positive and negative examples, and hence any pre-chosen test set should consist of words from both classes. However, we also need a proper discussion of the proportion of positive and negative examples in the test set, as the most likely scenario in any given text is "no change".

[19] It is, of course, always possible to augment this "gold set" with new examples. Gold truth creation, though, is extremely costly both in time and money: Schlechtweg et al. (2020) report a total cost of EUR 20,000 (1,000 hours) for 37 English words, 48 in German, 40 in Latin, and 31 in Swedish.

(ii) Evaluating the output of the algorithm allows us to evaluate the performance of a method "in the wild" and truly study its behaviour. Unfortunately, this evaluation strategy requires new evaluation with each change either to the method or the data, as there potentially can be a completely new set of words to evaluate each time. The words to evaluate can be chosen on the basis of a predetermined measure of change (e.g., largest / smallest cosine angle between two consecutive time periods, i.e., the words that changed the most or least), or a set of randomly chosen words. Once a set of words is determined, the evaluation of each word is done in the same manner as for the pre-chosen test set.

The *accuracy* of the evaluation, regardless of strategy chosen, depends on the way we determine if and how a word has changed. The ground-truth must be constructed from the data (corpus) on which the methods are trained because existing dictionaries might list changes seen in the *language*, that might not be present in the corpus, or vice versa. Requiring a method to find change that is not present in the underlying text, or considering detected changes as false because they are not present in a general-purpose dictionary, both lead to artificially low performance of the method. When (manually) creating ground-truth data for evaluation, sample sentences from the dataset should be read and taken into consideration, thus grounding the change in the dataset.

4.1.2 Simulated LSC

Obtaining ground-truth data for LSC is a difficult task as it requires skilled annotators and takes time to produce. The problem is exacerbated as the time depth of the language change phenomena increases and the languages at hand become rarer. This fact leads to a further requirement: expert annotators. The notion of "expert annotator" is problematic when judging senses in the past. Previous studies (e.g. Schlechtweg et al. 2018) note that historical linguists tend to have better inter-annotator agreement between themselves than with "untrained" native speakers – hinting at the fact that this is a skill that can be honed. The difficulty of engaging sufficiently many expert annotators is also a theoretical argument in favour of synthetic evaluation frameworks as a complement. In addition, some types of LSC are less frequent than others,[20] therefore requiring large amounts of text to be annotated in order to find enough samples. To alleviate these problems, simulating LSC in existing corpora has been suggested.

[20] Assuming that semantic change is power-law distributed, like most linguistic phenomena.

Simulating LSC is based on a decades-old procedure of inducing polysemy to evaluate word sense disambiguation systems (Gale et al. 1992, Schütze 1998). In this approach two or more words (e.g., *chair* and *sky*) are collapsed into a single word form (*chairsky*), thus conflating their meanings and creating a pseudo polysemous word (the original pair is removed from the lexicon). From another perspective, if this procedure unfolds through time (i.e., a word either gained or lost senses), then it can be considered to simulate LSC via changes to the number of senses of words. Indeed, this approach has been used extensively to simulate LSC (Cook & Stevenson 2010, Kulkarni et al. 2015, Rosenfeld & Erk 2018, Shoemark et al. 2019). However, the LSC that is induced in this way is rather synthetic, because it collapses unrelated words into a single word form, as opposed to the general view that finds the different senses to be semantically related (Fillmore & Atkins 2000). In order to provide a more accurate LSC simulation, Dubossarsky et al. (2019) accounted for the similarity of candidate words prior to their collapse, both creating related pairs (e.g., *chair* and *stool*) that better reflect true polysemy, and comparing the pair with the original approach of unrelated pairs (*chair* and *sky*). Schlechtweg & Schulte im Walde (2020) use SemCor, a sense-tagged corpus of English, to control for the specific senses each word has at each time point, thus providing an even more ecological model for simulated LSC.

The simulated approach to LSC has the potential to circumvent any bottleneck related to the need for annotators, and thus reduces costs. In addition, with careful planning, it should be possible to simulate any desirable type of LSC, regardless of its rarity in natural texts. As an added bonus, and certainly of interest to lexicographers, such an evaluation allows us to compute recall. In this scenario, recall would be proportional to the number of changed words that a given method can find. Using a synthetic change dataset is currently the only realistic scenario for determining the recall of our models and therefore, detecting how much change a method is able to capture. At the same time, it is hard to argue against the legitimate concern that these LSCs are artificial, and as such may not be the optimal way to evaluate detection by computational models. Certainly, synthetic change datasets are not optimal to study the natural linguistic phenomenon of semantic change, at least before we have a full understanding of the large-scale phenomena that we wish to study at which point we might no longer be in need for synthetic datasets. However, without the considerable effort to annotate full datasets, we are bound to use synthetic change evaluation sets – despite the inherent limitation described above. As a result, an important factor for future research becomes the creation of synthetic datasets that reflect the complex and varying nature of real language and real semantic change.

We stipulate that simulated datasets should be used alongside ground-truth testing, both with respect to pre-chosen test sets, as well as evaluating the output, to properly evaluate the ability of any method to detect LSC.

4.2 Quantifying meaning and meaning change

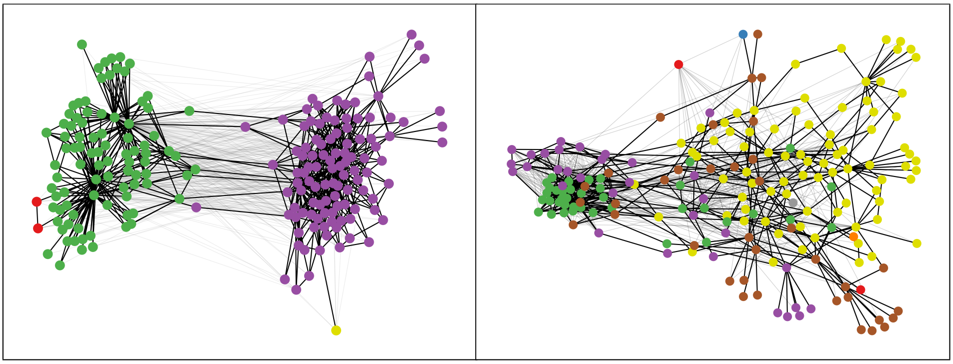

Figure 11.1: Word usage graphs of German *zersetzen* (left) and *Abgesang* (right).

To provide high-quality, ground-truth data for LSC where word meaning and change is grounded in a given corpus, we must perform manual annotation. However, first, we need to choose the relevant level of meaning so that we can quantify meaning distinctions and the change for a word based on the annotation. Recently, SemEval-2020 Task 1 (Schlechtweg et al. 2020) implemented a *binary* and a *graded* notion of LSC (Schlechtweg & Schulte im Walde 2020) in the shared task, which was partly adopted by a follow-up task on Italian (Basile et al. 2020). The annotation used typical meaning distinctions from historical linguistics (Blank 1997). Although the authors avoided the use of discrete word senses in the annotation by using graded semantic relatedness judgements (Erk et al. 2013, Schlechtweg et al. 2018), they grouped word uses post-hoc into hard clusters and interpreted all uses in a cluster as having the same sense. While this discrete view can work well in practice for some words (Hovy et al. 2006), it is inadequate for others (Kilgarriff 1997, McCarthy et al. 2016). In order to see this, consider Figure 11.1, showing the annotated and clustered uses for two words from the SemEval dataset: the uses of the word *zersetzen* on the left can clearly be partitioned into two main clusters, while the ones of *Abgesang* on the right have a less clearly clusterable structure.

A graded notion of meaning and change can be used to avoid having to cluster cases like the latter, though it is still unclear what the practical applications could be for LSC without discrete senses. The advantage of discrete word senses is that, despite their inadequacy for certain words, they are a widely used concept and also build a bridge to historical linguistics (Blank 1997, Blank & Koch 1999). This bridge is an important one, because the most straightforward application of LSC detection methods is for historical linguistics or lexicography. Nonetheless, there might be many task-specific definitions of LSC that could do without sense distinctions, and the issue is an interesting avenue for future work.

5 Related fields and applications

The field of LSC has close ties with two types of disciplines: those that study (i) *language*, and those that study (ii) *human activities*. In this section, we shed light on prominent work in these fields without claiming to be exhaustive, and discuss the potential of interactions with these fields.

5.1 Studying language

A great deal of existing work has gone into the study of language. Lexicography benefits a great deal from semantic representation in time, with works by, among others, Lau et al. (2012), Falk et al. (2014), Fišer & Ljubešić (2018), Klosa & Lüngen (2018), and Torres-Rivera & Torres-Moreno (2020). In this strand, methods for LSC can prove efficient at updating historical dictionaries: by using LSC approaches on large-scale corpora, it becomes possible to verify, at the very least, whether a sense was actually used *before* its current date in the dictionary. Senses cannot be post-dated, on the other hand; their absence from a corpus does not necessarily mean they did not exist elsewhere. Lexicographers can ideally use these methods to generate *candidates* for semantic change which would then be manually checked. They could also use sense-frequency data to paint the prominence of a word's senses through time, or even incorporate a quantified measure of similarity between senses of the same word – features that could also be incorporated in contemporary dictionaries.

Another strand, despite most work focusing solely on English, concerns language in general. In the past few years, there have been several attempts at testing hypotheses for laws of change which were proposed more than a century ago, or devising new ones. Xu & Kemp (2015) focus on two incompatible hypotheses: Bréal (1897)'s LAW OF DIFFERENTIATION (where near-synonyms are set to diverge across time) and Stern (1921)'s LAW OF PARALLEL CHANGE (where words sharing related meanings tend to move semantically in the same way). They showed quantitatively for English, in the Google Books corpus, that Stern's law of parallel change seems to be more rooted in evidence than Bréal's law of differentiation. Dubossarsky et al. (2015) ground their LAW OF PROTOTYPICALITY on the hypothesis of Geeraerts (1997) that a word's relation to the core prototypical meaning of its semantic category is crucial with respect to diachronic semantic change, and show using English data that prototypicality is negatively correlated with semantic change. Eger & Mehler (2016) postulate and show that semantic change tends to behave linearly in English, German and Latin. Perhaps the best-known example of such work within NLP, and often the only one cited, are the two laws of Hamilton et al. (2016b): CONFORMITY (stating that frequency is negatively correlated with semantic change), and INNOVATION (hypothesising that polysemy is positively correlated with semantic change).

Interestingly, since the NLP work above derives from observations that are replicable, quantitative, somewhat evidentiary, and not from a limited set of examples as was the case in the early non-computational days of semantic change research, previous laws elicited from quantitative investigations can be revisited. Such was the aim of Dubossarsky et al. (2017). They show that three previous laws (the law of prototypicality of Dubossarsky et al. 2015 and the laws of innovation and conformity by Hamilton et al. 2016b) are a byproduct of a confounding variable in the data, namely frequency, and are thus refuted. The paper calls for more stringent standards of proof when articulating new laws – in other words, robust evaluation.

As regards future work, we envision the field of LSC moving towards better use of linguistic knowledge. Traditional semantics and semantic change research is deeply rooted in theories that can now be computationally operationalized. Additionally, advances in computational typology and cross-lingual methods allow language change to be modelled for several similar languages at the same time (as started by Uban et al. 2019 and Frossard et al. 2020, for example), and to take into account theories of language contact. Other linguistic features can also be taken into account, and we hope to see more work going beyond "simple" lexical semantics.[21] The overview and discussions in this chapter have primarily targeted semantic change, often referred to as semasiological change in linguistic literature, while onomasiological change relates to different words used for the same concepts at different points in time. This general concept is often referred to as lexical replacement (Tahmasebi et al. 2018). Future work should attempt to resolve onomasiological and semasiological change in an iterative manner to ensure coherency in our models.

[21] An excellent example of this move forward can be seen in Fonteyn (2020), for example.

5.2 Studying human society

Along with the study of language itself, NLP techniques can be repurposed to serve different goals. With NLP methods maturing and technical solutions being made available to virtually anyone,[22] theories in other fields can be tested quantitatively. Quite obviously, since language is humans' best tool of communication, advanced techniques that tackle language are useful in many other fields, where they are often applied as-is, and sometimes modified to serve a different purpose. What is often disregarded in NLP, however, is what we need from those tangential disciplines in order to arrive at reliable models. One obvious answer to this question pertains to data resources. Those who are working on semantic change computation are heavily dependent on the data at their disposal, and should pay more attention to the type, diversity and quality of data they are working with, as discussed in Section 2.

In this subsection we focus on a few examples of how related fields have borrowed methods from LSC by using some examples from the literature, and attempt to give broad avenues for a continued mutualistic relationship between LSC and those fields.

A great deal of past human knowledge that has survived is stored in texts. Historical research[23] is arguably a large beneficiary of proper semantic representations of words in time: an often voiced critique in historical scholarship relates to chronological inconsistencies – anachronisms (Syrjämäki 2011). As reported by Zosa et al. (2020), Hobsbawm (2011) stated that "the most usual ideological abuse of history is based on anachronism rather than lies". This fact leads to many historians trying to "see things their [the people of the past's] way" (Skinner 2002). Somewhat similarly, Koselleck (2010) underlines the "veto right of the sources". However, for one to use the sources properly, they need to be understood correctly, and proper modelling of a word's semantics across time can definitely help historians interpret past events. Furthermore, the "concepts as factors and indicators of historical change" of Koselleck (2004: 80) highlights the importance of language as a window on the past. There is a growing body of work with quantitative diachronic text mining (such as word embeddings and (dynamic) topic models) within humanities research which clearly benefits from NLP methods, but can similarly inform LSC. For example, Heuser (2017)[24] studies the difference

[22] For example, extremely large-scale pretrained models are shared on platforms such as Hugging Face (https://huggingface.co/models) allowing anyone to download and use them with limited hardware; while efficient libraries such as gensim (Řehůřek & Sojka 2010) make the training of type embeddings possible on personal laptops.

[23] For clarity's sake, we do not differentiate between "historical research", "digital history", "computational history", and "digital humanities". For a broader discussion about field-naming in the (digital) humanities, refer to Piotrowski (2020).

[24] See https://twitter.com/quadrismegistus/status/846105045238112256 for a visualisation.

between abstract and concrete words in different literary subgenres. Similarly, Björck and co-authors[25] study the Swedish word for 'market', *marknad*, and describe a change in abstractness through time: from a physical market (as a noun), to more and more abstract notions (such as 'labour market') and even to the point where the noun is used as a modifier (e.g. 'market economy'). These observations teach us not only about the word itself, but also about the world. If words such as *table* or *car* are relatively straightforward to define and probably easier to model (see e.g. Reilly & Desai 2017 who show that concrete words tend to have denser semantic neighbourhoods than abstract words), what lessons can we learn from such work when representing abstract concepts? LSC methods should strive to include such key information in its methods.

Claiming that current LSC methods can "solve historical research"[26] and provide definitive answers to long-studied phenomena would be, at best, extremely misleading. Indeed, while LSC methods can model a word's sense(s) across time, humanists (or political scientists, for that matter) can be described as studying *concepts*. An emerging or evolving concept, almost by definition, will not be constrained to a single word. Rather, methods will probably have to be adapted to study a cluster of words[27] – either manually chosen (Kenter et al. 2015, Recchia et al. 2016), or selected in a more data-driven way (Tahmasebi 2013, Hengchen et al. to appear). These clusters will be the basis for historical contextualisation and interpretation. The same ad-hoc adaptation is to be found in political science: a recent example of NLP methods making their way in (quantitative) political science is the work of Rodman (2020) where the author fits both an LDA model on more than a century of newspapers as well as a supervised topic model – using 400 hand-annotated documents by several annotators with a high inter-annotator agreement – so as to produce a gold standard to evaluate diachronic word embeddings, with the final aim of studying the evolution of *concepts* such as 'gender' and 'race'. Similar work is undertaken by Indukaev (2021), who studies modernisation in Russia and convincingly describes the benefits and limitations of topic models and word embeddings for such a study.

While extremely promising, our current methods fail to serve related fields that would benefit greatly from them: as of now, most LSC approaches simply model words, and not concepts – again underlining the need for task-specific meaning tackled in Section 3.

[25] Presentation by Henrik Björck, Claes Ohlsson, and Leif Runefelt given at the Workshop on automatic detection of language change 2018 co-located with SLTC 2018, Stockholm. For more details, see Ohlsson (2020).

[26] Or any field concerned with diachronic textual data.

[27] These clusters of words are related to what linguists call lexical fields, a term that in our experience is not widely used in other disciplines.

6 Conclusions

In this chapter, we have outlined the existing challenges in reconstructing the semantic change evident in large diachronic corpora.

Currently, as was made obvious in Section 2, the field suffers from several limitations when it comes to data. Indeed, we believe that future work should strive to use, and produce, high-quality text in many languages and different genres. This point is crucial: if we are to detect semantic change, our data needs to have certain precise qualities, as well as well-defined metadata. It is thus difficult for LSC researchers to rely on data created for other NLP fields. As a plea to the larger community, we count on the field not to make the mistake of assuming that the available textual data is representative of the language at hand. We further hope that in the future, meaning can be modelled by using not only text, but also multi-modal data.

Modelling is notoriously difficult, but, to paraphrase Box (1976), models being inherently wrong does not ineluctably make them useless. A crucial component to the useful modelling of meaning and of change outlined in Section 3 is the definition of what meaning is. Whether we draw from the large body of work that exists in traditional semantics research or start from scratch with a new definition of what *computational* meaning is, we hope researchers in our field can come together and agree on *what* is and should be modelled. Only with shared, solid models of meaning can the field move forward with the complexity, possibly intractable, of modelling meaning change. A word's semantics have changed – but how?

Echoing the complexity of modelling information from data is the consistency needed in the evaluation of a model's output. Section 4 makes the point that without a homogeneous, somewhat large-scale evaluation framework across languages such as the one proposed in Schlechtweg et al. (2020), researchers cannot confidently rely on conclusions from previous work to move forward. Since ground-truth creation is expensive both in time and money and is ineluctably limited to a single corpus, we encourage the community to pay attention to synthetic evaluation techniques which have the potential to circumvent cost, evaluate different types of semantic change, and tackle different temporal granularities. Our field is rich in methods but in dire need of comparable results. This can be partially solved with robust, thorough, and shared evaluation practices.

Being able to model and detect different types of semantic change is important in LSC, and also in related disciplines such as lexicography and historical linguistics. The history of ideas, and any area concerned with the diachronic study of textual data, would greatly benefit from our methods – if they are robust. In addition, we believe that there is potential for a mutualistic relationship with those parallel fields not only contributing theory or domain expertise but also echoing the need for the proper modelling of words, senses, and types of change.

Author contributions

All authors contributed equally, and the ordering is determined in a round robin fashion.

Acknowledgements

We thank our reviewers for their very helpful comments. This chapter benefited from insightful comments from Lars Borin. This work has been funded in part by the project *Towards computational lexical semantic change detection* supported by the Swedish Research Council (2019–2022; dnr 2018-01184), and *Nationella språkbanken* (the Swedish National Language Bank), jointly funded by the Swedish Research Council (2018–2024; dnr 2017-00626) and the ten partner institutions. Dominik Schlechtweg was supported by the Konrad Adenauer Foundation and the CRETA Centre funded by the German Ministry for Education and Research (BMBF) during the writing of this chapter. Haim Dubossarsky is supported by the Blavatnik Postdoctoral Fellowship Programme. The authors wish to thank the Department of Swedish at the University of Gothenburg for providing financial support for language editing services.

Abbreviations

ACL	Association for Computational Linguistics
COHA	corpus of historical American English
LDA	latent Dirichlet allocation
LSC	lexical semantic change
NLP	natural language processing
OCR	Optical Character Recognition
SLTC	Swedish Language Technology Conference
WSD	word sense disambiguation

References

Aristotle. 1898. *Poetics.* (English translation by Ingram Bywater). Oxford: The Clarendon Press.

Bamler, Robert & Stephan Mandt. 2017. Dynamic word embeddings. In Doina Precup & Yee Whye Teh (eds.), *Proceedings of the 34th international conference on machine learning* (Proceedings of Machine Learning Research 70), 380–389. Sydney: PMLR.

Basile, Pierpaolo, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti & Rossella Varvara. 2020. Overview of the EVALITA 2020 diachronic lexical semantics (DIACR-ita) task. In Valerio Basile, Danilo Croce, Maria Di Maro & Lucia C. Passaro (eds.), *Proceedings of the 7th evaluation campaign of natural language processing and speech tools for Italian (EVALITA 2020).*

Bender, Emily M. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology* 6(3). 1–26.

Bender, Emily M. & Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of ACL 2020*, 5185–5198. Online: ACL. DOI: 10.18653/v1/2020.acl-main.463.

Blank, Andreas. 1997. *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen.* Tübingen: Niemeyer.

Blank, Andreas. 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* (Cognitive Linguistics Research 13), 61–90.

Blank, Andreas & Peter Koch. 1999. *Historical semantics and cognition.* Berlin: Walter de Gruyter.

Bloomfield, Leonard. 1933. *Language.* New York: Henry Holt.

Bowern, Claire. 2019. Semantic change and semantic stability: Variation is key. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 48–55.

Box, George EP. 1976. Science and statistics. *Journal of the American Statistical Association* 71(356). 791–799.

Boydens, Isabelle. 1999. *Informatique, normes et temps.* Bruxelles: Bruylant.

Bréal, Michel. 1897. *Essai de sémantique.* Paris: Hachette.

Bruni, Elia, Gemma Boleda, Marco Baroni & Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th annual meeting of the Association for Computational Linguistics (Volume 1: Long papers)*, 136–145.

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

Cook, Paul & S. Stevenson. 2010. Automatically identifying changes in the semantic orientation of words. In N. C. C. Chair, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (eds.), *Proceedings of LREC 2010*. Valletta: ELRA.

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

Denny, Matthew J. & Arthur Spirling. 2018. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis* 26(2). 168–189.

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

Dubossarsky, Haim, Yulia Tsvetkov, Chris Dyer & Eitan Grossman. 2015. A bottom up approach to category mapping and meaning change. In *Proceedings of NetWordS 2015* (CEUR Workshop Proceedings 1347), 66–70. Pisa.

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

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

Erk, Katrin, Diana McCarthy & Nicholas Gaylord. 2013. Measuring word meaning in context. *Computational Linguistics* 39(3). 511–554.

Falk, Ingrid, Delphine Bernhard & Christophe Gérard. 2014. From non word to new word: Automatically identifying neologisms in French newspapers. In *Proceedings of LREC 2014*, 4337–4344. Reykjavik: ELRA.

Fillmore, Charles J. & Beryl T. S. Atkins. 2000. Describing polysemy: The case of 'crawl'. *Polysemy: Theoretical and computational approaches* 91. 110.

Fišer, Darja & Nikola Ljubešić. 2018. Distributional modelling for semantic shift detection. *International Journal of Lexicography* 32(2). 1–21. DOI: 10.1093/ijl/ecy011.

Fonteyn, Lauren. 2020. What about grammar? Using BERT embeddings to explore functional-semantic shifts of semi-lexical and grammatical constructions. *CEUR Workshop Proceedings* 1613. http://ceur-ws.org/Vol-2723/short15.pdf.

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

Gale, William A., K. W. Church & D. Yarowsky. 1992. Work on statistical methods for word sense disambiguation. In *Working Notes of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, 23–25. http://www.aaai.org/Papers/Symposia/Fall/1992/FS-92-04/FS92-04-008.pdf.

Geeraerts, Dirk. 1997. *Diachronic prototype semantics: A contribution to historical lexicology*. Oxford: Oxford University Press.

Giulianelli, Mario, Marco Del Tredici & Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of ACL 2020*, 3960–3973. Online: ACL. DOI: 10.18653/v1/2020.acl-main.365.

Gulordava, Kristina & Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, 67–71. Edinburgh: ACL. https://www.aclweb.org/anthology/W11-2508.

Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016a. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *Proceedings of EMNLP 2016*, 2116–2121. Austin: ACL. DOI: 10.18653/v1/D16-1229.

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

Harris, Zellig S. 1954. Distributional structure. *Word* 10(2-3). 146–162.

Hätty, Anna, Dominik Schlechtweg & Sabine Schulte im Walde. 2019. SURel: A gold Standard for incorporating meaning shifts into term extraction. In *Proceedings of the 8th Joint Conference on Lexical and Computational Semantics*, 1–8. Minneapolis.

Hellrich, Johannes & Udo Hahn. 2016. Bad company: Neighborhoods in neural embedding spaces considered harmful. In *Proceedings of COLING 2016: Technical papers*, 2785–2796. Osaka: ACL. https://www.aclweb.org/anthology/C16-1262.

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

Hengchen, Simon, Ruben Ros & Jani Marjanen. 2019. A data-driven approach to the changing vocabulary of the *nation* in English, Dutch, Swedish and Finnish newspapers, 1750–1950. In *Proceedings of the Digital Humanities (DH) conference 2019*.

Hengchen, Simon, Ruben Ros, Jani Marjanen & Mikko Tolonen. to appear. A data-driven approach to studying changing vocabularies in historical newspaper collections. *Digital Scholarship in the Humanities*. DOI: 10.1093/llc/fqab032.

Heuser, Ryan James. 2017. Word vectors in the eighteenth century. In *Book of abstracts of the 2017 Digital Humanities conference (DH2017)*. Montréal.

Hobsbawm, Eric. 2011. *On History*. London: Hachette UK.

Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw & Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of NAACL-HLT: Short papers* (NAACL-Short '06), 57–60. New York: ACL.

Hu, Renfen, Shen Li & Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of ACL 2019*, 3899–3908. Florence: ACL. DOI: 10.18653/v1/P19-1379.

Indukaev, Andrey. 2021. Studying ideational change in Russian politics with topic models and word embeddings. In Daria Gritsenko, Mariëlle Wijermars & Mikhail Kopotev (eds.), *Palgrave handbook of Digital Russia Studies*, 443–465. Basingstoke: Palgrave Macmillan.

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

Kaiser, Jens, Dominik Schlechtweg, Sean Papay & Sabine Schulte im Walde. 2020. IMS at SemEval-2020 Task 1: How low can you go? Dimensionality in lexical semantic change detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*. Barcelona: ACL.

Kenter, Tom, Melvin Wevers, Pim Huijnen & Maarten De Rijke. 2015. Ad hoc monitoring of vocabulary shifts over time. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, 1191–1200.

Kilgarriff, Adam. 1997. "I don't believe in word senses". *Computers and the Humanities* 31(2). 91–113.

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

Klosa, Annette & Harald Lüngen. 2018. New German Words: Detection and Description. In *Proceedings of the XVIII EURALEX international congress lexicography in global contexts, 17–21 July 2018, Ljubljana*, 559–569. Ljubliana: Ljubljana University Press.

Koplenig, Alexander. 2016. *Analyzing lexical change in diachronic corpora.* Universität Mannheim. (Doctoral dissertation). http://nbn-resolving.de/urn:nbn:de:bsz:mh39-48905.

Koselleck, Reinhart. 2004. *Futures past: On the semantics of historical time.* New York, NY: Columbia University Press.

Koselleck, Reinhart. 2010. *Vom Sinn und Unsinn der Geschichte: Aufsätze und Vorträge aus vier Jahrzehnten.* Carsten Dutt (ed.). 1st edn. Berlin: Suhrkamp.

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

Kutuzov, Andrey, Erik Velldal & Lilja Øvrelid. 2017. Temporal dynamics of semantic relations in word embeddings: An application to predicting armed conflict participants. In *Proceedings of EMNLP 2017*, 1824–1829. Copenhagen: ACL. DOI: 10.18653/v1/D17-1194.

Lau, Jey Han, Paul Cook, Diana McCarthy, David Newman & Timothy Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of EACL 2012*, 591–601. Avignon: ACL. https://www.aclweb.org/anthology/E12-1060.

Magué, Jean-Philippe. 2005. *Changements sémantiques et cognition: Différentes méthodes pour différentes échelles temporelles.* Lyon: Université Lumière. (Doctoral dissertation).

McCarthy, Diana, Maria Apidianaki & Katrin Erk. 2016. Word sense clustering and clusterability. *Computational Linguistics* 42(2). 245–275.

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

Meillet, Antoine. 1905. Comment les mots changent de sens. *Année Sociologique.*

Michel, Jean-Baptiste, 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. *Science* 331. 176–182.

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

Ohlsson, Claes. 2020. Market language over time. Combining corpus linguistics and historical discourse analysis in a study of market in Swedish press texts.

In Joacim Hansson & Jonas Svensson (eds.), *Doing digital humanities: Concepts, approaches, cases*, vol. 1, 199–218. Växjö: Linnaeus University.

Pechenick, Eitan Adam, Christopher M. Danforth & Peter Sheridan Dodds. 2015. Characterizing the Google Books corpus: Strong limits to inferences of sociocultural and linguistic evolution. *PLOS ONE* 10(10). e0137041.

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

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

Piotrowski, Michael. 2012. Natural language processing for historical texts. *Synthesis lectures on human language technologies* 5(2). 1–157.

Piotrowski, Michael. 2020. Ain't no way around it: Why we need to be clear about what we mean by "digital humanities". *SocArXiv*. DOI: 10.31235/osf.io/d2kb6.

Ponti, Edoardo Maria, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova & Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics* 45(3). 559–601. DOI: 10.1162/coli_a_00357.

Prescott, Andrew. 2018. Searching for Dr. Johnson: The digitisation of the Burney Newspaper Collection. In Siv Gøril Brandtzæg, Paul Goring & Christine Watson (eds.), *Travelling chronicles: News and newspapers from the early modern period to the eighteenth century*, 51–71. Leiden: Brill.

Recchia, Gabriel, Ewan Jones, Paul Nulty, John Regan & Peter de Bolla. 2016. Tracing shifting conceptual vocabularies through time. In *European knowledge acquisition workshop*, 19–28. Springer.

Řehůřek, Radim & Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*, 45–50. http://is.muni.cz/publication/884893/en. Valletta: ELRA.

Reilly, Megan & Rutvik H. Desai. 2017. Effects of semantic neighborhood density in abstract and concrete words. *Cognition* 169. 46–53. DOI: 10.1016/j.cognition.2017.08.004.

Reisig, Karl. 1839. *Vorlesungen über lateinische Sprachwissenschaft*. Leipzig: Lehnhold.

Rodda, Martina A., Marco S.G. Senaldi & Alessandro Lenci. 2017. Panta rei: Tracking semantic change with distributional semantics in Ancient Greek. *Italian Journal of Computational Linguistics* 3(1). 11–24. https://arpi.unipi.it/handle/11568/891899.

Rodman, Emma. 2020. A timely intervention: Tracking the changing meanings of political concepts with word vectors. *Political Analysis* 28(1). 87–111.

Rosenfeld, Alex & Katrin Erk. 2018. Deep neural models of semantic shift. In *Proceedings of NAACL-HLT 2018: Volume 1 (Long papers)*, 474–484. New Orleans: ACL. DOI: 10.18653/v1/N18-1044.

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

Säily, Tanja. 2016. Sociolinguistic variation in morphological productivity in eighteenth-century English. *Corpus Linguistics and Linguistic Theory* 12(1). 129–151.

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

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

Schlechtweg, Dominik & Sabine Schulte im Walde. 2020. Simulating lexical semantic change from sense-annotated data. In C. Cuskley, M. Flaherty, H. Little, Luke McCrohon, A. Ravignani & T. Verhoef (eds.), *The Evolution of Language: Proceedings of the 13th International Conference (EVOLANGXIII)*.

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

Schofield, Alexandra, Måns Magnusson & David Mimno. 2017. Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of EACL 2017 (Volume 2: Short papers)*, 432–436.

Schofield, Alexandra & David Mimno. 2016. Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the ACL* 4. 287–300. DOI: 10.1162/tacl_a_00099.

Schofield, Alexandra, Laure Thompson & David Mimno. 2017. Quantifying the effects of text duplication on semantic models. In *Proceedings of EMNLP 2017*, 2737–2747. Copenhagen: ACL. DOI: 10.18653/v1/D17-1290.

Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics* 24(1). 97–123.

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

Skinner, Quentin. 2002. *Visions of politics. Vol. 1, Regarding method.* Cambridge: Cambridge University Press.

Stern, Nils Gustaf. 1921. *Swift, swiftly and their synonyms. A contribution to semantic analysis and theory.* Gothenburg: Göteborgs universitet. (Doctoral dissertation).

Syrjämäki, Sami. 2011. *Sins of a historian: Perspectives on the problem of anachronism.* Tampere: Tampere University Press.

Tahmasebi, Nina. 2013. *Models and algorithms for automatic detection of language evolution.* Gottfried Wilhelm Leibniz Universität Hannover. (Doctoral dissertation). http://edok01.tib.uni-hannover.de/edoks/e01dh13/771705034.pdf.

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

Tahmasebi, Nina, Lars Borin, Adam Jatowt & Yang Xu (eds.). 2019. *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change.* Florence: ACL. https://www.aclweb.org/anthology/W19-4700.

Tahmasebi, Nina & Simon Hengchen. 2019. The strengths and pitfalls of large-scale text mining for literary studies. *Samlaren: Tidskrift för svensk litteraturvetenskaplig forskning* 140. 198–227.

Tahmasebi, Nina, Kai Niklas, Gideon Zenz & Thomas Risse. 2013. On the applicability of word sense discrimination on 201 years of modern English. *International Journal on Digital Libraries* 13. 135–153. DOI: 10.1007/s00799-013-0105-8.

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

Tangherlini, Timothy R. & Peter Leonard. 2013. Trawling in the sea of the great unread: Sub-corpus topic modeling and humanities research. *Poetics* 41(6). 725–749.

Thompson, Laure & David Mimno. 2018. Authorless topic models: Biasing models away from known structure. In *Proceedings of COLING 2018*, 3903–3914. Santa Fe: ACL. https://www.aclweb.org/anthology/C18-1329.

Torres-Rivera, Andrés & Juan-Manuel Torres-Moreno. 2020. Detecting new word meanings: A comparison of word embedding models in Spanish. *arXiv preprint 2001.05285*.

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

Tsakalidis, Adam & Maria Liakata. 2020. Sequential modelling of the evolution of word representations for semantic change detection. In *Proceedings of EMNLP 2020*, 8485–8497. Online: ACL. https://www.aclweb.org/anthology/2020.emnlp-main.682.

Uban, Ana, Alina Maria Ciobanu & Liviu P. Dinu. 2019. Studying laws of semantic divergence across languages using cognate sets. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 161–166. Florence: ACL. DOI: 10.18653/v1/W19-4720.

Voigt, Rob, Nicholas P. Camp, Vinodkumar Prabhakaran, William L. Hamilton, Rebecca C. Hetey, Camilla M. Griffiths, David Jurgens, Dan Jurafsky & Jennifer L. Eberhardt. 2017. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences* 114(25). 6521–6526. DOI: 10.1073/pnas.1702413114.

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

Xu, Yang & Charles Kemp. 2015. A computational evaluation of two laws of semantic change. In *Proceedings of the 37th annual meeting of the Cognitive Science Society, CogSci 2015*. Pasadena.

Zosa, Elaine, Simon Hengchen, Jani Marjanen, Lidia Pivovarova & Mikko Tolonen. 2020. Disappearing discourses: Avoiding anachronisms and teleology with data-driven methods in studying digital newspaper collections. In *Book of abstracts of the 2020 digital Humanities in the Nordic countries (DHN) conference*.