Computational Models of Semantic Change

Nina Tahmasebi
University of Gothenburg, Sweden

Andrey Kutuzov
University of Oslo, Norway

Haim Dubossarsky
Queen Mary University of London, UK

Mario Giulianelli
University College London, UK

The Wiley Blackwell Companion to Diachronic and Historical Linguistics. Edited by Adam Ledgeway, Edith Aldridge, Anne Breitbarth, Katalin É. Kiss, Joseph Salmons, and Alexandra Simonenko.
 Published 2026 by John Wiley & Sons, Inc.
DOI: 10.1002/9781119898023.wbcdl140

1 Introduction

The computational study of semantic change is a field that has emerged over the past two decades. Previously, studies of semantic change were limited to the manual study of single words or small, typically thematic, groups of words. Examples include the study of vocabulary relating to color or temperature, and of the English verb 'do' over the course of a century. These studies were carried out either by close reading of textual data, supplemented by a large pinch of intuition, or by trials in which subjects were shown images or color samples and were asked to name them (Vejdemo 2017).

This type of qualitative study of word meaning and meaning change allows researchers to conduct in-depth analysis of reasons and causes, and to hypothesize both from and to the examples studied. It has laid the basis for proposed schema of semantic change; Bréal (1899) and Stern (1931) proposed classifications of types of change, including broadening and narrowing of senses, amelioration and pejoration, and metaphorization. These general processes could, in theory, cover all possible types of semantic change. The strength of these qualitative studies lies in the large amount of theory and domain knowledge that has been applied to interpret and generalize the data. However, as is common with qualitative studies, they are limited both as regards the number of texts analyzed and the number of words studied simultaneously.

Quantitative studies of semantic change only became possible with the introduction of large-scale diachronic corpora in digital format, for example, the Helsinki Corpus (Rissanen and Ihalainen 1991), the British National Corpus (BNC) (Leech et al. 1992), and the Google Books Ngrams corpus (Michel et al. 2011), and the development of powerful computational tools. The first computational studies in this field, like Sagi, Kaufmann, and Clark (2009), largely replicated the methods of qualitative studies by following a few words in detail across longer time spans. However, what was new was that (i) meaning was defined computationally and (ii) the comparison over time (of computationally defined meaning) was quantified. More studies followed, using different data (types of text and time spans), and different computational models to define meaning (Lau et al. 2012; Tahmasebi 2013). Despite employing larger datasets, the majority of these early studies remained limited in the number of words that were analyzed in depth.

Eventually, larger-scale studies were conducted. Mitra et al. (2014) studied tens of thousands of candidate novel senses and evaluated a small portion of these manually. Gulordava and Baroni (2011) studied 10,000 words in the Google Books Ngrams corpus consisting of about 150 million n-grams (see more on this corpus later in the chapter), though only a random sample of 100 words was manually evaluated. These larger-scale studies gave a hint of the strengths of computational modeling of semantic

change. Dubossarsky et al. (2015) introduced the first 'law of semantic change' (the law of prototypicality) on the scale of a whole vocabulary, trying to computationally measure qualitatively proposed laws. Soon more laws followed (Xu and Kemp 2015 and Hamilton, Leskovec, and Jurafsky 2016b). However, these 'laws' exposed a weakness in computational modeling on the basis of large-scale digital texts, namely that *these models were detecting inherent properties of the datasets themselves rather than detecting true semantic change*. Dubossarsky, Weinshall, and Grossman (2017) showed that if we shuffle the texts around so that each time period contains texts from all other time periods, thus eliminating or significantly reducing the presence of semantic change, the methods still detected change. Therefore, what these models were detecting was a bias inherent to most diachronic corpora, namely, the growing amount of texts over time leads to an artificial change, rather than true change in meaning. This leads us to the question of what exactly computational methods are measuring. While manual qualitative analysis is limited in regard to the number of instances and the volume of texts that can be used, larger-scale computational analysis requires a thorough understanding of *what* the models are capturing. Essentially, one type of bias can easily be replaced by another.

Nonetheless, computational study of semantic change (also called 'lexical semantic change detection' or LSCD) has opened vast new arenas of research: one can now study the processes of semantic change across languages and different genres, over specific times, and in specific contexts. We can revisit the manually devised classifications and laws to find empirical evidence of the limits to semantic change theories. Perhaps, the soundest path forward is not an either-or but rather a synergy between the computational study of semantic change and the qualitative, theory-based approach.

1.1 Aspects of semantic change

Many aspects of semantic change can be modeled computationally. Varying combinations of these aspects shape and constrain relevant research questions and suitable methods. Here we discuss four main axes of variation along which most computational studies of semantic change can be positioned: the nature of change, the source of change, the level of analysis, and the timescale. This list is not exhaustive, but it provides a structure to this body of research and allows us to appreciate its breadth, especially in terms of questions and motivations. Appropriate methods will be the topic of Section 2.

1.1.1 The nature of change

Describing the nature of change, that is, the ways in which the meaning of a word changes over time, has been a topic of study in modern diachronic linguistics since the late nineteenth century.

In his *Prinzipen der Sprachgeschichte*, Hermann Paul distinguishes three main categories of semantic change: meaning specialization (or narrowing), generalization (or broadening), and transfer (Paul 1880) – the latter subsumes what would later be called metaphorization and metonymization. More categories were identified in the following decades. In his popular taxonomy, Leonard Bloomfield enumerates nine types of change (Bloomfield 1933), including metaphor, metonymy, synecdoche, hyperbole, and meiosis, as well as pejoration and amelioration. Stephen Ullmann has proposed a more recent taxonomy that recognizes similarity and contiguity as two fundamental types of association underlying meaning change, as well as designation and meaning as the

Table 1 Common types of semantic change (OE = Old English).

Type	Word	Old meaning	New meaning
Broadening	*bird*	a young bird	> any kind of bird
Narrowing	*meat*	food	> edible flesh
Amelioration	OE *cniht*	servant boy	> knight
Pejoration	*awful*	inspiring wonder	> terribly bad
Metaphorization	*broadcast*	cast seeds	> transmit reports

spaces in which associations can be formed. Associations in the space of designation give rise to cases of folk etymology and ellipsis (Ullmann 1962). Andreas Blank (1997) provides an extensive historical overview.

Some categories appear across the taxonomies of different linguists. Among these are narrowing and its opposite, broadening. For example, a word can undergo a *broadening* of its meaning, like 'bird', the meaning of which has changed from a *young bird* to *any kind of bird*. Another commonly studied category of semantic change is *metaphorization*, as in the case of 'broadcast', the meaning of which has changed from *spreading seeds in the field* to *casting (or spreading) media reports, or sending any media from a mobile device to a TV*. However, there is no consensus on a general taxonomy of semantic change types. For a list of common types of change that are generally in agreement between researchers, see Table 1.

1.1.2 The source of change

Computational studies vary with respect to which source of change they investigate. Each source can tell a story of semantic change from a unique perspective. For example, changes in the real world, the *external* environment, including events, technological advances, and societal change, result in changes in word usage such as that captured by the change in meaning of the word *broadcast*. Language users parsimoniously accommodate these external changes by adding new meanings to existing words. Meaning change can also be determined by factors that are *internal* to a language, such as the degree of polysemy of a word and its degree of prototypicality (i.e., its representativeness of a certain semantic category). Such factors are hypothesized to be correlated with the word's potential for semantic change (Geeraerts 1997; Dubossarsky et al. 2015; Hamilton, Leskovec, and Jurafsky 2016b).

External pressures can be of different kinds. While sociolinguists may be interested in the *sociocultural factors* affecting, for example, whether speakers use *coach* to refer to a bus or to a carriage, psycholinguists may be more interested in the *cognitive factors* that lead speakers to assign the same word form the unrelated – or at best distantly related – meaning of sports instructor or trainer. Both groups of linguists may use computational systems of semantic change detection to analyze the meaning trajectories of the word *coach* (e.g., the study by Giulianelli, Tredici, and Fernández 2020). Some may describe or explain the change as a function of social pressures and will focus on, say, the relationship between community structure and change (Noble et al. 2021). Others may focus on cognitive pressures, for example, explaining English adjective extension in terms of the cognitive mechanism of chaining (Grewal and Xu 2021). Researchers can thus offer complementary but independent accounts of semantic change. The effects of different sources of change are of course intertwined, and accounts that attempt to combine social and cognitive explanations offer a more holistic and more faithful

perspective on semantic change (Brochhagen et al. 2023; Hawkins et al. 2023). Such accounts are, unfortunately, still rare.

1.1.3 The level of analysis

We further suggest dividing the analysis of semantic change into five levels, ranging from detecting that the meaning of a word has changed over time to providing a causal explanation of its diachronic meaning trajectory. We summarize this view in Table 2.

The first level of analysis determines whether a word has indeed undergone meaning change over time. The subsequent levels of analysis employ the same categories as Table 1. These levels were proposed to describe change in concise and more informative ways, because it is rarely sufficient to merely report which words have changed their meaning.

On closer inspection, however, it seems that these categories are actually different in nature. Broadening seems to be a purely *descriptive* category, while metaphorization suggests a *mechanism of change*, that is, it classifies semantic change according to the process by which the new meaning emerged. We therefore view these two as separate levels of analysis of change. Each category tackles change at an increasing level of complexity, and each provides independent accounts for the change that may complement each other (e.g., metaphorization can result in broadening of a word's scope of meaning). The mechanistic account at the third level of analysis puts us in a better position to understand semantic change (relative to lower-level accounts). However, this still does not provide a complete causal account of why the change has happened. We therefore propose to view causal accounts of semantic change as the fourth and fifth levels of analysis. It is only at these levels that laws of semantic change become relevant.

Causal accounts of change in other branches of linguistics where change is also prevalent, such as morphology, phonology, and grammaticalization, predated the discussion of causality in semantic change. It is therefore not surprising to see that the first attempts to articulate laws of semantic change resembled, and perhaps were inspired by, the same rules that were proposed in other branches. Thus some argued that word frequency plays a role in driving language change (Bybee 2006), while others favored polysemy as a driver of semantic change (Traugott and Dasher 2001). Here too, we must distinguish between two types of causal accounts. The first one is internal, and is based on linguistic factors that drive (or correlate with) semantic change, like frequency or polysemy. The second type of account points to external factors, that is, cognitive and communicative factors that reside outside the linguistic system (see Section 1.1.2).

A priori, we attribute higher importance to external accounts as we deem them more robust. Internal accounts can be susceptible to circular reasoning, where the explanation relies heavily on the internal system being explained. External accounts

Table 2 Levels of analysis.

Level	Type	Description
1	Detection	Determining a word changed its meaning
2	Descriptive categorization	Classify the change by specific type
3	Mechanistic explanation	Classify how the change came to be
4	Causal (internal)	Linguistic variables that predict change
5	Causal (external)	Non-linguistic variables that predict change

are more credible because they introduce independent factors that provide validation and avoid circularity.

Computational models of semantic change still fall short when it comes to most of these levels of analysis. In fact, current models focus mainly on the entrance level of complexity, that is, on the detection of semantic change events. Only recently have serious attempts been made to tackle the second level of analysis. The source of the problem is the design limitations of computational models when it comes to addressing causal analysis of semantic change (see further discussion in Section 2.2).

Nonetheless, a few laws of semantic change have been proposed based on large-scale analysis (Dubossarsky et al. 2015 and Hamilton, Leskovec, and Jurafsky 2016b). However, as it turned out, much of this research was premature and the proposed laws were later severely criticized due to reliability issues and a lack of scientific rigor (Dubossarsky, Weinshall, and Grossman 2017; Dubossarsky et al. 2019). As of now, computational models of semantic change still need to rise to the challenge of the different levels of analysis (e.g., Keidar et al. 2022), though some initial steps have been taken (Cassotti, de Pascale, and Tahmasebi 2024).

1.1.4 The timescale

Semantic change can be investigated on different timescales. Change can occur over the course of a few weeks or months. Sometimes this change is temporary and mainly affects the frequency and relative proportions of different word meanings. In January 2020, for example, journalists writing about COVID-19 were using the word *strain* almost exclusively to indicate the variant of a virus or bacterium; by April, more than six out of ten occurrences in news articles referred to *strain* as an excessive demand on resources – financial, infrastructural, or healthcare systems (Montariol, Martinc, and Pivovarova 2021). Over the same months, the word *virus* underwent a fast and extreme process of specialization in almost everyone's vocabulary, with usages of *virus* specifically denoting a new referent, SARS-CoV-2 (or COVID-19), rather than the general concept of a replicating infectious agent. Such changes are often referred to as examples of *short-term meaning change*.

The most felicitous cases of specialization and generalization of word meaning succeed in taking a more stable place in the lexicon. For example, usage of the word *mouse* to refer to a pointing device is attested to have gradually become more prominent throughout the second half of the twentieth century (Wijaya and Yeniterzi 2011), and this meaning is likely here to stay for at least another few decades. Looking further into the future, however, when personal computers may be replaced by devices with more innovative pointing mechanisms, it is easy to imagine this use of *mouse* becoming obsolete. Other types of referents never go out of fashion and, for these, it is sometimes possible to observe trajectories of meaning change spanning multiple centuries. For example, in the fourteenth century the word *girl*, which now refers exclusively to a child of female gender, referred to a young person or child of either sex. Such instances of long-term change are a topic of great interest in diachronic research on classical languages such as Latin and Greek, where the corpora under analysis can span two millennia (McGillivray and Kilgarriff 2013; Vatri and McGillivray 2018). The Greek word *mus*, for instance, is cohabited by three main senses (*mouse*, *muscle*, and *mussel*) the relative prominence of which oscillates between the eighth century BCE and the fifth century CE (Perrone et al. 2021).

When diachronic analyses focus on one or a few words, and attempt to precisely track how their multiple senses coexist over long time periods (Nowak 2019, for example,

studying the evolution of the signal word *tempus* time in Latin from the second century BCE to the sixth century CE), it is important to use corpora with extensive coverage of multiple genres as well as methods that can account for how different genres select specific word senses (Perrone et al. 2019). In sum, diachronic studies can investigate meaning change trajectories over varying timescales, taking any position on a continuum of *temporal granularity* that ranges from days and weeks to decades and centuries.

The aspects presented in this section can hopefully help the reader make sense of the vast and heterogeneous landscape of computational studies of semantic change. While we have presented them as independent aspects, they are naturally interconnected. For example, studies that look at language internal factors are more likely to focus on large timescales like centuries, while analyses of short-term meaning change typically focus on societal factors.

1.2 Scenarios

Computational methods for semantic change detection can be applied in a variety of research fields to answer different kinds of research questions, ranging from questions that pertain to language itself to questions that relate primarily to our societies and cultures as these are documented in text. Most clearly, methods for detecting large-scale semantic change are applicable in lexicography and historical linguistics. However, the study of conceptual change is relevant to several fields of research in the broader humanities and social sciences where comparison of concepts before and after events or time periods can be very important. These concepts may be concrete and easily defined or they may be abstractly defined cultural and societal phenomena present in text. In this section, we will provide example scenarios and outline some research directions in which semantic change methodology can play an important role.

1.2.1 Lexicographic and linguistic research

Lexicographic work to create or update dictionaries currently involves a large degree of manual effort, particularly when it comes to detecting subtle semantic changes. Lexicographers often have to rely on chance encounters, manual crawling of news, social media, and literature, or targeted work within certain areas of the vocabulary that are prone to change, like technical vocabulary or vocabulary pertaining to sensitive descriptors of matters such as sexuality, partnership, immigrants, and disabilities. Computational models for semantic change have the potential to contribute extensively to the task of semantic change *discovery* (Zamora-Reina, Bravo-Marquez, and Schlechtweg 2022). They can scan the full vocabulary, from the last update of the dictionary until the present, to identify words that have experienced a change in meaning. The definition of meaning here is straightforward lexical meaning. Depending on the lexicographic tradition, different granularities of meaning can also be considered.

In historical linguistic research, the object of study is often broader and can pertain both to semantic fields and to lexical classes. The interest is often multilingual, examining how changes in one language compare to or affect another. Computational methods for semantic change have been employed in this field with various degrees of success in terms of completeness and ability to answer the full research question. While much remains to be done, we can learn from the directions that have already been taken.

Early work to compare the change degrees of verbs, nouns, and adjectives in general was performed by Dubossarsky, Weinshall, and Grossman (2016), while Hamilton et al.

(2016) showed that many *evaluative* adjectives in English have completely switched their sentiment during the last 150 years (probably due to their emotional load). In this work, 'evaluative' adjectives were defined as those that describe object qualities from the subjective point of view of the speakers, expressing their opinions about the object being described.

Another example of a large-scale study performed with computational methodology is that of Rodina et al. (2019), who studied the *intensity* of diachronic semantic change in evaluative adjectives. They set out to determine whether evaluative adjectives are more prone to semantic change than other types of adjectives – that is, whether the shift in their meaning is more probable and occurs faster. It is known that amelioration and pejoration occur on a massive scale, and examples of evaluative words changing their polarity can be found in Traugott and Dasher (2001). But is their change really stronger and faster than that of other adjectives? The answer to this question is of obvious importance to a lexicographer, since it helps to better prioritize efforts when updating dictionaries and other resources.

Rodina et al. (2019) employed six different methods of quantifying semantic change to analyze data in three languages (English, Norwegian, and Russian), over a span of five decades (from the 1960s to 2000s). Frequency-controlled experiments showed that, depending on the particular method, evaluative adjectives either did not differ from other types of adjectives in terms of semantic change or were actually less prone to it. Thus, in spite of many well-known examples of semantically changing evaluative adjectives, it seems that these processes are not particularly characteristic of this specific type of word, at least when considering relatively short-term time spans. However, when language data are observed over a longer time, computational methods start to capture a slow and consistent movement of evaluative adjectives away from their original meaning.

In this scenario, the aim of the study is *descriptive*: the researcher wants to test which groups of words change in a more intensive way than other groups, without trying to find the causes of such behavior. The temporal granularity is *decades*, which can be considered midway between long-term timescales (centuries) and short-term timescales (years or months).

1.2.2 Conceptual change

The study of semantic change is at its essence a study of word meaning displayed in text. However, in text we encode much more than the meaning of individual words. Therefore, by studying the change of information in texts, we can easily study changing perceptions or changes in our cultures and societies. This second scenario relates to conceptual change and is relevant to all text-based research fields, and in particular to the broader humanities and social sciences.

The evolution of concepts is a long-standing topic within philosophy, history, and linguistics. Here, our definition of meaning is much broader than lexical meaning and includes connotational meaning, or culturally and socially defined concepts. The changing valence of a word like *girl* from any young person to a virgin, unmarried young female to any younger girl is an indication of changes in our views and our societal values, and has less to do with lexical meaning change. Concepts can be defined as single words or as combinations of words that together form a concept that cannot be described by any one word. Examining the latter obviously requires using different methods to computationally define meaning.

If these concepts are tracked over longer periods of time, typically, we first need to detect *onomasiological* change (see Section 3.2). For example, if we are interested in studying how women have been viewed over time in Swedish, we must be aware that the word for woman has had both different spellings and different word forms: *kona* → *qwinna* → *qvinna* → *kvinna*.

Conceptual studies may have a historical or a modern perspective. For example, Hengchen, Ros, and Marjanen (2019) studied how -isms (liberalism, socialism, and conservatism) relate to ideological language and contributed to our understanding of the development of political language over time. Combining conceptual history and economic history, the Market Language Project (Ohlsson, Wåhlstrand Skärström, and Björck 2022) studies how the concept of a market has become increasingly abstract over time, so that the *market* as an agent in our everyday lives is far from the marketplaces that the word referred to in the Middle Ages. Vylomova, Murphy, and Haslam (2019) study semantic changes in concepts related to harm in psychology and investigate the hypothesis that certain concepts such as *addiction*, *bullying*, *harassment*, *prejudice*, and *trauma* have broadened over the last four decades. Tripodi et al. (2019) track the evolution of anti-Semitic bias in the religious, ethical, economic, sociopolitical, racial, and conspiratorial domains. The results show a trend of growing anti-Semitism, starting in the mid-1980s in France.

In a study by Sommerauer and Fokkens (2019), computational models of semantic change were used to study the cultural evolution of racism, working from the hypothesis that the concept has changed from being perceived as related to visual biological attributes such as skin color and is now subject to cultural interpretations regarding differences between groups of people. When carrying out conceptual studies, it is very important to translate the informal hypotheses into a set of precise verifiable hypotheses on how concept evolution is reflected in diachronic changes in language use. One way to do this is by compiling lists of *core concept terms* and *related concept terms*. Examples of core concept terms for racism are *race, culture, racial,* and *cultural,* while examples of related concept terms are *skin color, nationality,* and *language*. These concept terms can, in turn, have core concepts like *whites, blacks, Jews,* and *Arabs,* or include terms that are more at the margin of the core meaning of interest, such as *superior* and *inferior* or *genetics* and *values*. Together, these constitute a conceptual network (Betti and Van den Berg 2014). The evolution of *racism* can then be tested by examining whether relations between the elements in this network change over time. More precisely, the researcher hypothesizes that related concept terms such as *skin color, superior,* and *inferior* will have moved away from the core concepts, while terms such as *national, language,* and *values* will have moved closer to the core.

Methods to derive such semantic spaces from text corpora will be described later in this chapter. For now it suffices to know that these representations are typically numerical vectors that position words in a high-dimensional semantic space, such that words which are similar in meaning occupy the same region of the space and semantically unrelated words are distant from each other. An alternative solution is to use similar computational models to test whether loose analogies of the form '*racism* is to *genetics* at the beginning of the 20th century as *racism* is to *value* at the end of the 20th century' hold according to the textual data under scrutiny (Orlikowski, Hartung, and Cimiano 2018).

Scenarios like this tend to have some common characteristics: the research goal is descriptive. There is no particular focus on the sources of change, but they are assumed to be external (i.e., sociocultural factors); the analysis mode is semantic change tracking;

and the study is onomasiological in approach, in that it is anchored to a concept, *racism*, and analyzes networks of word forms associated with this abstract concept. The temporal granularity is coarse, as this example scenario addresses changes spanning a whole century. However, similar research questions can be asked about faster trajectories of concept evolution.

This example operates on single words, rather than a cluster of words to form a concept, and the same distributional spaces that are used to capture lexical meaning. Future research is needed to develop methods that can allow us to define meaning either from a purely connotational viewpoint or with different cultural or social concepts, like values, opinions, or emotions.

2 Current computational approaches

The past 15 years have seen a surge of interest in and research on developing computational models for understanding meaning and, by extension, meaning change. These models vary in terms of architectural design, computational complexity, and scope. However, they all share a common foundation, namely, the distributional hypothesis (DH). In the following section, we will present the DH and its different implementations, as well as the relation between methods and the data being modeled.

2.1 The distributional hypothesis

The DH suggests that meaning can be inferred by systematically analyzing word contexts, as words with similar meanings tend to occur in similar contexts. For example, encountering the word *cat* in sentences like *a cat is meowing* or *the cat drinks milk* allows us to make a reasonable guess about its meaning.

Introduced independently in the 1950s by Zellig Harris and John F. Firth, the DH posits that similarity of meaning equals similarity of usage. Despite its early inception, the DH remained relatively isolated. In our view, its primary contribution may have been its stance in the nature versus nurture debate on language acquisition (Brunila and LaViolette 2022). Unlike generative linguistics, the DH asserts that meaning arises solely from language usage, disregarding innate cognitive concepts.

The reason for its limited application beyond its academic niche was primarily practical. The DH requires extensive amounts of text for processing, making its use in modeling meaning and its experimental validation as a theory challenging or infeasible. However, the current availability of vast digitized textual resources coupled with recent advances in computational power and text processing algorithms has revolutionized the landscape. These developments have facilitated the emergence of larger and more sophisticated DH-based computational models, broadening their adoption in research and real-world applications.

The significant advances in DH-based models, also known as distributional semantic models (DSM), have made them the prevailing approach in modern natural language processing (NLP) and machine learning methods. They power numerous AI applications. However, their success and ease of deployment have resulted in their overshadowing of alternative methods that employ dictionaries, lexicons, ontologies, or knowledge graphs to model and represent meaning. The overwhelming focus on a single semantic theory centered on the DH means that other theories which could offer more elaborate and potentially more accurate models of meaning are being neglected.

2.1.1 Roots and validation of the DH

The roots of the DH can be traced back to various research disciplines that have contributed to its development. One significant influence has been Ferdinand de Saussure's structuralist perspective in linguistics. Saussure emphasized the relationship between signs and their associative networks within a language system, asserting that meaning arises from the interplay of signs. In a way, the DH operationalizes the structuralist viewpoint. For example, in *a cat is meowing*, the word *cat* can be interchanged with pet-like animals. However, it cannot be replaced by inanimate nouns such as *ball* or *iPad*, nor with words from a different syntactic category like *anxious* or *shining*. This structuralist viewpoint aligns with the fundamental principles of DH, as it recognizes the importance of analyzing the distribution and context of linguistic elements to infer meaning.

The DH finds additional support in psycholinguistic research on language learning. The phenomenon of entrenchment, which refers to the formation of associative connections between words and their contexts in our cognitive system (Schmid 2017), provides empirical evidence for the DH. Through repeated exposure to words in specific usage contexts, our cognitive system develops strong associations, enabling us to predict and anticipate certain words based on their co-occurrence patterns (Ramscar et al. 2014; Baayen et al. 2017). This psycholinguistic perspective strengthens the notion that meaning is derived from language usage and context, as postulated by the DH.

The DH has also been extensively validated on an empirical basis by evaluating the distributional representations generated by computational models. At their core, all DSMs represent the meaning of a word (or a sentence) using a vector of real numbers (a vector is simply an array, or long list, of numbers). In line with Harris (1954) who equated similarity of meaning to similarity of usage, it follows that the more similar two vectors are, the more similar their meaning is, which is the crux of the validation of the DH. Word vectors are evaluated using intrinsic evaluation tasks, such as word similarity or word analogy tasks, where the performance is measured against human judgment (Finkelstein et al. 2002; Hill, Reichart, and Korhonen 2015). The success of these evaluations in aligning with human intuition provides empirical evidence supporting the DH.

2.2 Data-driven approaches and methods

In this section, we provide a high-level description of modern computational approaches to the detection of semantic changes. These methods are by and large data driven – that is, they rely on large collections of raw textual data (corpora). One of the main reasons for this is that for most languages and time periods, we lack manually engineered historical ontologies or historical dictionaries. Sometimes these resources exist, but are hidden behind a paywall. Thus, the field needed unsupervised and statistical methods to infer the diachronic meaning of words from textual data.

Since it is impossible to collect the full corpus of a language, any study deals with a specific *sample* of language data. Researchers thus have to rely on the assumption that the sample they are studying is representative of the language, or aspects thereof, that they are interested in. This assumption of sample representativeness is extremely important: if a sample does not contain the necessary signal, the signal obviously cannot be found. However, the corpora that we work with do not need to be representative of languages as a whole: many research fields are interested in studying the

works of a single author, a small social group, or other limited contexts. More on this is given in Section 2.3.

The choice of data is followed by the choice of an appropriate method to analyze the data. This choice revolves around questions of how to *represent* semantic entities in a machine-readable way (often as numeric vectors) and which method can best detect the signals that are of interest.

2.2.1 Overview of meaning representations

The spectrum of existing data-driven ('distributional') representations of meaning is very rich. Among many other distinctions, one is particularly important: if the meaning is expressed by vectors, these vectors can be sparse and explicit (with interpretable components) or dense and distributed (with non-interpretable components). In modern NLP, trained dense vectors, also known as 'word embeddings', are generally preferred (Baroni, Dinu, and Kruszewski 2014).

Before the era of word embeddings, other representations were also employed. These can still be useful for some purposes, in particular for scenarios where less data are available. In principle, a word can be represented by its corpus frequency only: changes in raw word frequencies can be used to trace semantic shifts or other kinds of linguistic change. For examples of such work, see Juola (2003), Hilpert and Gries (2009), Michel et al. (2011), Lijffijt, Säily, and Nevalainen (2012), and Bochkarev, Solovyev, and Wichmann (2014), or Choi and Varian (2012) for frequency analysis of words in web search queries. Frequency-based methods can be useful in detecting the emergence of neologisms (Ryskina et al. 2020). The method can be as simple as calculating the absolute or normalized difference between target word frequencies in two time-specific corpora.

However, if one wants to trace the semantic change of an existing word form, using raw frequency differences has obvious limitations. Semantic shifts are not necessarily accompanied by strong changes in word frequency (or the connection may be very subtle and indirect). Since words belong to different frequency tiers, and absolute frequency values are not distributed across the vocabulary uniformly, it is difficult to find a robust method to calculate frequency differences between diachronic corpora. Nowadays, as a rule, frequency is used only as the simplest possible baseline for semantic change detection systems (Schlechtweg et al. 2020).

Around 2009, it was proposed that vector-based distributional methods could reliably detect semantic shifts that are not manifested through frequency change or simple collocates change. The pioneering work of Jurgens and Stevens (2009) presented an insightful conceptualization of a sequence of distributional representations changing over time: it is effectively a *Word*×*Semantic Vector*×*Time* tensor, in the sense that each word possesses a set of semantic vectors for each time span of interest. The more different the time-specific vectors, the higher the degree of semantic change between the corresponding time bins.

This concept paved the way for quantitative comparison of not only the synchronic meaning of words but also of different stages in the development of word meaning over time. It still remains the foundation of the whole field of semantic change modeling.

Jurgens and Stevens (2009) employed the Random Indexing (RI) algorithm (Kanerva, Kristofersson, and Holst 2000) to create word vectors from a training corpus, while Sagi, Kaufmann, and Clark (2009) turned to latent semantic analysis (Deerwester et al. 1990). Both methods are now outdated, but they already worked with dense vectors. Technically, the only difference between these representations and modern

word embeddings is that they were not trained using language modeling. However, back then the 'word embedding revolution' was still several years ahead. Two years later Gulordava and Baroni (2011) still used explicit representations consisting of sparse word co-occurrence matrices weighted by local mutual information. A similar approach was taken by Tahmasebi et al. (2012), who traced the evolution of named entities.

The diversity of the methods used increased over time, with graph approaches gaining popularity. For example, Mitra et al. (2014) analyzed the clusters of a word similarity graph in sub-corpora corresponding to different time periods. Their distributional model consisted of lexical nodes in graphs connected by weighted edges. The weights corresponded to the number of shared most salient syntactic dependency contexts, where saliency was determined by co-occurrence counts scaled by mutual information. Importantly, they were able to detect not only the mere fact of a semantic shift, but also its type: the birth of a new sense, splitting an old sense into several new ones, or merging of several senses into one. Other examples of graph-based approaches are Tahmasebi (2013) and Tahmasebi and Risse (2017a), who tracked individual sense changes (word-sense evolution) on the basis of the curvature clustering algorithm.

Another vein of research employed *topic modeling* approaches (where topics are interpreted as senses). A prominent example is Lau et al. (2012), who applied latent Dirichlet allocation (LDA) in conjunction with a nonparametric hierarchical Dirichlet process. Senses were naturally mapped to automatically inferred corpus topics, so that the distribution of word senses corresponded to the topic probabilities.

Parametric distributional models that produce *word embeddings* using machine learning (ML) methods from raw corpora have been used to model semantic change since 2014. Word embeddings are dense continuous vector representations of lexical semantics trained in an iterative unsupervised fashion with the objective of minimizing loss on the language modeling objective. Language modeling is here understood as the task of predicting the next word in a sentence, given the previous words (or, sometimes, predicting a masked-out word given the words surrounding it). Thus, ultimately the source of the signal is still word co-occurrence counts in the training corpus, but the resulting representations are more efficient and convenient to work with than previous methods (Baroni, Dinu, and Kruszewski 2014). The latest development is the use of *contextualized* or *token-based* models like ELMo (Peters et al. 2018) or BERT (Devlin et al. 2019). Such models yield different embeddings for one and the same word form (token) in different contexts.

Kutuzov, Pivovarova, and Giulianelli (2021) proposed 'grammatical profiling', that is, comparing diachronic distributions of *morphological* and *syntactic* features to obtain insights about *semantic* change. They found that this method yielded convincing results, with morphological and syntactic categories being complementary (i.e., combining them improves semantic change detection performance). Importantly, the predictions derived from grammatical profiling are interpretable and are thus suitable for linguistic studies that require qualitative explanations. Later, Giulianelli, Kutuzov, and Pivovarova (2022) showed that *ensembling* grammatical profiles with contextualized embeddings improves the performance of semantic change detection for most benchmarking datasets and languages.

2.2.2 *Type-based models*

The earliest learned distribution models are *type-based* models or *static* word embedding models. Typical examples are architectures like word2vec (Mikolov et al. 2013)

and FastText (Bojanowski et al. 2017). Following their widespread adoption in NLP in general, they quickly became dominant representations for the analysis of semantic change as well. This is clearly evidenced by the results of the first SemEval shared task in unsupervised lexical semantic change detection (Schlechtweg et al. 2020): 18 of the 21 participants (including all the winners) used either static or contextualized word embeddings. In the simplest form, these involve producing two vector representations for a word 'X' in two different time periods (C_1 and C_2) and then computing the cosine similarity between these two vectors. If the representations are very similar (the cosine similarity is high, close to 1), then the meaning of 'X' did not change. Otherwise, a shift has occurred.

Historically, the work of Kim et al. (2014) has been seminal in the sense that they were arguably the first to employ word embedding models to trace diachronic semantic shifts. They used word2vec. They also introduced the incremental or chronological training approach to leverage new properties of embeddings. Kim et al. (2014) successfully identified semantic shifts in widely used examples like the English word *cell* (at the beginning of the twenty-first century). However, the limitations of such a method were already clear. It cannot be used to determine the nature of the shift (e.g., narrowing or widening, amelioration or pejoration). In addition, because of the one-vector-per-word paradigm, the method offers 'average' representations of the different senses of a word. Dominant senses will be much more influential in determining where in distributional space the vector should be placed. For example, for the word *rock* the music sense is currently much more prominent than the stone sense, and also more prominent than verb usages like *to rock*. A vector for *rock* trained in a modern corpus will likely be placed in a neighborhood exclusively representing words related to music and none related to stone. As a consequence, not only do these methods have limitations in terms of detecting the nature of a shift, but they also cannot detect changes to any *smaller* sense, nor represent the less dominant senses individually.

Comparison over time. One issue that is specific to type-based word embeddings is the problem of making embedding spaces comparable (sometimes called 'the problem of alignment') so that vectors from different time periods can be compared. Comparing vector representations across embedding models that have been separately trained is not straightforward, even if the vocabulary is essentially the same. The problem stems from the stochastic nature of word embeddings. The most popular remedy is to *align* different vector spaces by somehow making the vectors comparable (i.e., the vectors for semantically similar stable words trained on C_1 should yield high cosine similarity to those trained on C_2). Kulkarni et al. (2015) suggested fitting the models into one vector space, using linear transformations to preserve the general space structure. If we are given two independently trained embedding matrices A and B with a significant shared vocabulary (e.g., trained on two diachronic corpora), we can find an orthogonal linear transformation T such that it projects A to B while minimizing the squared loss, for example, by using the orthogonal Procrustes method (Gower and Dijksterhaus 2004).

After A is projected to the B vector space, cosine similarities between their vectors become meaningful and can be used as indicators of semantic change. Using direct alignment with orthogonal projections is easy and straightforward. This approach is sometimes criticized for its self-contradictory objective (it attempts to project each word to itself, even in the presence of a shift) and for instability with respect to different embedding spaces (Gonen et al. 2020). However, it is still very efficient in semantic

change detection with type-based word embeddings, as shown by Shoemark et al. (2019) and others.

2.2.3 *Contextualized models*

Since 2018, there has been a paradigm shift: contextualized embedding architectures like ELMo (Peters et al. 2018) or BERT (Devlin et al. 2019) now offer an entirely new approach to the problem of making diachronic embeddings comparable, and allow more detailed studies. Unlike the 'static' embedding models, the contextualized models work with *token embeddings*, that is, context-dependent representations of words. After training, the model can be used to infer embeddings for each occurrence of a target word in time-specific corpora. These token embeddings will be similar for tokens used in similar contexts and different for tokens used in different contexts. This makes it possible to formulate various ways of estimating the difference of token embeddings between two or more time-specific corpora, without the need for alignment; the embeddings are produced by one and the same language model and are thus comparable by design.

One major challenge with contextualized embeddings is that because they are more fine-grained, they result in much larger models (i.e., LLMs – large language models). This has implications for the amount of training data needed; typically, several orders of magnitude more data is required to train a contextualized model than a static embedding model. These models are therefore typically *pretrained* on a very large corpus before being released online. Pretraining is usually computationally expensive, requiring time and optimized hardware. It is assumed that in the process of pretraining, an LLM acquires foundational statistical knowledge about language. *Fine-tuning* it for a specific task then becomes much less expensive. For research in the field of semantic change detection, the use of LLMs means that diachronic corpora are used to *infer* word representations rather than to *train* them. Currently, the overwhelming majority of LLMs are based on the transformer neural architecture. Foundational models actively used in LSCD include BERT (Devlin et al. 2019) and XLM-R (Conneau et al. 2020).

Although the first studies applying contextualized language models to semantic change modeling only started to appear in 2019–2020 (Hu, Li, and Liang 2019; Giulianelli, Tredici, and Fernández 2020; Martinc, Novak, and Pollak 2020; Martinc et al. 2020; Kutuzov and Giulianelli 2020), the use of pretrained contextual models has already become the de facto standard. At the RuShiftEval shared task on LSCD for Russian (Kutuzov and Pivovarova, 2021a), the leaderboard was dominated by contextualized models. See Periti and Montanelli (2024) for an extensive survey. Below we briefly introduce the most important methods.

Change detection using token-based embeddings. While change detection using type-based models is straightforward (once the embedding spaces have been aligned) because each word corresponds to a single vector in each time period, change detection using token embeddings is different because a word has as many token embeddings as the word has occurrences in the corpus corresponding to a time period. Using LLMs, we thus end up with two *usage matrices*, $\mathrm{U}_w^{t_1}$ and $\mathrm{U}_w^{t_2}$, corresponding to two time bins and containing token embeddings of a word w for each of the two time periods. Once the token embeddings are inferred (with whatever LLM chosen), it is necessary to choose a strategy to compare the represented time periods and compute a *change score*, indicating the degree of semantic change undergone by a word between them. There are a number of strategies for doing this.

One possible approach is clustering token embeddings into groups loosely corresponding to word senses. These can then either be treated as senses and directly compared over time, or be compared using their time-specific distributions (Martinc et al. 2020; Cuba Gyllensten et al. 2020; Giulianelli, Tredici, and Fernández 2020). Despite being promising in principle and offering a way forward to detect different change categories, clustering has not yet outperformed other, simpler methods. Further work is needed to evaluate different strategies for deriving sense-representations given token representations.

A straightforward method to avoid clustering is to average the token embeddings in a particular time bin, thus ending up with a single vector per time period, similar to the case with type-based embeddings. See for example Kutuzov and Giulianelli (2020) for details of the so-called PRT algorithm. However, averaging leads to loss of information. To avoid this, an alternative approach, the APD algorithm introduced by Giulianelli, Tredici, and Fernández (2020), leverages all token embeddings by accounting for the average distance between all possible pairs of token embeddings in $U_w^{t_1}$ and $U_w^{t_2}$ in the different time periods. High APD values indicate a higher degree of semantic change. Kutuzov, Velldal, and Øvrelid (2022) showed that averaging the PRT and APD estimates yields (on average) better predictions than either of the algorithms separately. This method is now known as PRT/APD or APD-PRT (Giulianelli, Kutuzov, and Pivovarova 2022).

Another interesting mode of using LLMs for semantic change modeling is presented by Homskiy and Arefyev (2022) and further developed by Cassotti et al. (2023). They employed XLM-R fine-tuned on a Word-in-Context (WiC) task, that is, a task determining whether a word w has the same meaning in two given contexts (in a way, replicating human annotations with a neural network). This model compares pairs of sentences containing w from two different time bins and produces probabilities of w being used in different senses for each pair. These probabilities are then used as distances in the APD algorithm.

Using contextualized models to represent meaning has great advantages for semantic change detection. It provides us with rich representations that can be used to differentiate senses of a word and advances our ability to move on to the further levels of analysis shown in Table 2. It is, however, still challenging to derive reasonable sense representations from token representations. In a broader sense, we have not yet made sufficient advances in the field of word sense discrimination. In addition, we are not yet utilizing all the information stored in an LLM: BERT, for example, has 12 layers of representations per word given a context and another 12 attention heads that can be explored, either individually or in combination. Initial explorations into layers has been done by Periti and Tahmasebi (2024), as well as Periti et al. (2024).

2.3 The interrelation between methods and data

The computational study of semantic change was sparked by the availability of digital textual corpora spanning longer time frames and the development of powerful computational tools that facilitated the implementation of the distributional hypothesis. The field was thus not born of an intentional push to use particular data and methods to study linguistic phenomena. As a result, the first decade of the LSCD field was devoted to randomly chosen datasets, and was a trial phase in terms of methods. As with all data-intensive research, different combinations of methods applied to data

provide us with different answers. If we set the data as, for example, a large newspaper corpus, and then apply different types of models to encode meaning, the end results will be radically different, depending on whether we use word sense induction methods, topic models, static word embeddings, or large pretrained contextualized embeddings. These differences relate both to which meanings we deem a word to have and to which words we deem to have changed meaning. If we instead set the method and change the data to which it is applied from a newspaper corpus to a collection of historical books, the outcome will again differ.

These properties mean that we are not primarily studying changing meanings across the language when we apply different methods to different datasets. Instead, we are only studying the combination of method and data. The important conclusion is that the insights we draw from different datasets using different methods cannot be used to evaluate language specific properties. Instead, what we are evaluating is corpora-specific properties. Analysis of short, information-intense newspaper articles will yield different insights from analysis of long books or five-word extracts of literary content. The information obtained may be complementary, but it can also be misleading or wrong.

We can think of several dimensions in which datasets differ. First, there is the matter of the time that the data spans, from Twitter corpora that cover one or a few years, to Latin texts that cover millennia. Second, there is the matter of the gap between the time periods being studied (commonly used gaps range from monthly data bins to century-long gaps and century-long bins). Third, we should consider the number of time bins: comparing two or three far-apart time periods is very different from studying hundreds of (daily, monthly, or yearly) time periods. Fourth, we need to consider the amount of data available in our dataset, from millions and billions of words, to thousands of words. The genre of text also matters. Certain datasets consist of genres that are information dense, like news articles that tend to provide us with most of the information in a few hundred words, whereas in books information can be spread over hundreds of pages of text. The differences also relate to the amount of data available for different time periods: for older stages of language, we typically have much less data. This relates both to digitized text and to printed text that could potentially be digitized.

To be able to draw insights that are language specific, we must use many different samples of texts that differ only in one dimension while applying the same methods to all of them. An important question that still remains open is the suitability of certain methods with respect to different datasets, depending on their properties.

2.4 Infrastructure

Lexical semantic change modeling requires substantial research infrastructure in terms of both data and algorithmic resources. In this section, we provide an overview of the available infrastructure resources.

2.4.1 Diachronic corpora

The most fundamental resources used in semantic change modeling are diachronic corpora, historical or modern, annotated with the time period when the texts were created. When modeling semantic change based on diachronic corpora, the types of generalizations that can be made are highly influenced by the properties of the textual data, where sources and temporal granularity play an important role. These diachronic corpora are

primarily used as sources to model semantic change, but can also be used to train or fine-tune language models.

The timescale (the granularity of the temporal dimension) is chosen before slicing the text collection into sub-corpora. Earlier works in semantic change dealt mainly with long-term semantic shifts (spanning decades or even centuries) since they are usually easier to trace. Early examples are Hilpert and Gries (2009) who studied the frequency developments of words in the TIME corpus[1] and Sagi, Kaufmann, and Clark (2009) who studied differences between Early Middle, Late Middle, and Early Modern English, using the Helsinki Corpus (Rissanen 1994).

Importantly, meaningful results in semantic change modeling with data-driven methods can be obtained only with large enough corpora, since a sufficient number of observations are required to yield statistically significant correlations as well as to train reasonable models (if machine learning is used). The increasing size and quality of digitized historical corpora in recent years is one of the reasons for the increased NLP research related to semantic change. A large role in the development of the field was played by the Google Books Ngrams corpus,[2] which led to the new data-driven discipline of 'culturomics', studying human culture through digital media (Michel et al, 2011). Mihalcea and Nastase (2012) used this corpus to detect differences in word usage and meaning across 50-year time spans, while a bit earlier Gulordava and Baroni (2011) compared word meanings in the 1960s and the 1990s. Unfortunately, Google Books Ngrams is inherently limited in that it does not contain full texts (it is only possible to download a maximum of five-word fragments). Nonetheless, for many cases, this corpus is adequate, and its use as a source of diachronic data continued in Mitra et al. (2014), who detected word sense changes over decades.

In much of the research cited below, time spans decreased in size and became more granular. In general, corpora with smaller time spans are useful for analyzing sociocultural semantic shifts, while corpora with longer spans are necessary for the study of linguistically motivated semantic shifts. As researchers are attempting to trace increasingly subtle cultural semantic shifts (often more relevant for practical tasks), the granularity of time spans is decreasing and the issue of *short-term semantic change* is receiving much attention. For example, Kim et al. (2014), Liao and Cheng (2016), and Del Tredici, Fernández, and Boleda (2019) analyzed yearly lexical changes.

In addition to the Google Ngrams corpus (with granularity of five years), Kulkarni et al. (2015) used Amazon Movie Reviews (with granularity of one year) and Twitter data (with granularity of one month). Their results indicated that computational methods to detect semantic shifts can be robustly applied to time spans of less than a decade. Since then, Twitter data has been a relatively popular choice for modeling short-term semantic change, with many datasets available, including the COVID-19 Twitter dataset containing around 14 billion word tokens (Banda et al. 2020). Another popular and publicly available corpus for short-term diachronic studies is the Signal Media Dataset (Corney et al. 2016) used, for example, by Kutuzov and Kuzmenko (2016).

Tahmasebi (2013) and Zhang et al. (2015) used the *New York Times* Annotated Corpus (Sandhaus 2008) with yearly sub-corpora, again managing to trace subtle semantic shifts. The same corpus was employed by Szymanski (2017) and to some extent by Yao et al. (2018), who crawled the *New York Times* website to obtain 27 yearly sub-corpora (from 1990 to 2016). Yao et al. (2018) captured semantic change with a granularity of years. For example, they observed that the nearest neighbors for the proper noun

Obama were moving from Barack Obama's pre-presidential life in 1990–2006 (e.g., *university, professor, civil*) to political terms in 2008–2016 (e.g., *president, campaign, government*), with similar trends observed for Donald Trump.

The inventory of diachronic corpora used in semantic change modeling was expanded by Jatowt and Duh (2014), who turned to the Corpus of Historical American English (COHA).[3] They used COHA as an additional source of data, with Google Ngrams being the main one. Hamilton, Leskovec, and Jurafsky (2016b) continued the usage of COHA along with the Google Ngrams corpus, and Eger and Mehler (2016) made the former their main data source (with a granularity of one decade). Cook et al. (2013) were the first to use two years of the English Gigaword news corpus (Parker et al. 2011), while Kutuzov, Velldal, and Øvrelid (2017) employed all its yearly slices in their analysis of cultural semantic change related to armed conflicts.

In Table 3 we list some popular English corpora that have been used for diachronic research with computational approaches. We give the sizes of the corpora in word tokens, but sheer size is not the only important property of a diachronic corpus. First, not all the corpora are publicly available: for example, the *New York Times* Annotated Corpus, COHA, and Gigaword are available for a fee only, while Google Books Ngrams does not provide any clear way to obtain the full corpus at all. Another aspect to consider is, of course, the time span covered by the corpus: the Helsinki Corpus might be small in comparison to Twitter or Gigaword, but if one is interested in Old English and Middle English, the latter corpora cannot be used. Finally, the domain composition of the corpus can be of paramount importance: diachronic shifts occurring in movie reviews can be very different from those occurring in news pieces.

Table 3 Commonly used English diachronic corpora.

Corpus	Size, words	Reference
Helsinki Corpus	10^6	Rissanen (1994)
New York Times Annotated Corpus	$\approx 2 \times 10^9$	Sandhaus (2008)
Google Books Ngrams	$\approx 100 \times 10^9$	Michel et al. (2011)
English Gigaword	$\approx 4 \times 10^9$	Parker et al. (2011)
Corp. of Hist. Amer. Engl. (COHA)	400×10^6	Davies (2012)
Amazon Movie Reviews	$\approx 9 \times 10^8$	McAuley and Leskovec (2013)
Twitter (also in other languages)	$\approx 14 \times 10^9$	Banda et al. (2020)

Table 3 is by no means exhaustive. English corpora also include the Corpus of Contemporary American English (COCA) and Project Gutenberg. Many other diachronic corpora are used for languages besides English, including the Deutsches Textarchiv, Berliner Zeitung, Neues Deutschland for German, the LatinISE for Latin, the Kubhist for Swedish, the Russian National Corpus and Lenta.ru dataset for Russian, and the NBDigital and Norsk Aviskorpus for Norwegian, and many others.[4]

2.4.2 Benchmarks

Diachronic corpora are needed not only as a source for the *detection* of semantic change but also as a source of *test sets* to evaluate such approaches. Evaluation test sets or benchmarks constitute another important infrastructure resource in computational semantic change modeling.

Works on language change originating in general linguistics like Traugott and Dasher (2001), Dobrushina and Daniel (2016) and others contain, as a rule, only a small number of handpicked examples. These smaller sets of words are not sufficient to properly evaluate automatic, unsupervised systems. The Dat-SemShift database (Zalizniak 2018) features more than 4000 semantic shifts across 800 languages. However, it is focused on cognitive proximities between pairs of linguistic meanings (with a limited set of predefined senses). In this paradigm, a semantic shift is just a case of extended polysemy. The Dat-SemShift database is extremely useful for identifying recurring cross-linguistic semantic shifts, but it remains an open issue whether it can be used to evaluate unsupervised semantic change detection systems.

Until recently, there were few standard test sets for semantic change modeling, and those that did exist were of varying quality and availability. For example, Gulordava and Baroni (2011) manually annotated a dataset of English words by the degree of their semantic change from the 1960s to the 1990s (the GEMS dataset). However, the authors did not make the GEMS publicly available. Even years after its publication, researchers still had to contact the authors personally to obtain the dataset.

Fortunately, the situation is improving. A prominent example is a package of test sets for English, German, Latin, and Swedish provided by Schlechtweg et al. (2020), accompanying the SemEval-2020 shared task 1. The datasets for Russian (Kutuzov and Pivovarova 2021b), Spanish (Zamora-Reina, Bravo-Marquez, and Schlechtweg 2022) and Norwegian (Kutuzov et al. 2022) follow the same approach. Most recently a similar dataset for Chinese was released (Chen et al. 2023). They are publicly available and manually annotated using a framework for the annotation of lexical semantic change called DURel or 'Diachronic usage relatedness' (Schlechtweg, Schulte im Walde, and Eckmann 2018). These datasets are based on the concept of so-called 'diachronic word usage graphs' (DWUGs).

DWUGs are produced using *graded* contextual word meaning annotation. In it, human annotators are shown two usages for the same word and asked to estimate their semantic similarity on a graded scale. Using these annotations, a graph is created for each target word, with word usages as nodes and annotators' judgments as weights on the edges. The graph is clustered into communities, assumed to represent (diachronic) senses. A numerical 'change score' is inferred from the changes that the distributions of these clusters undergo from one time period to another (Schlechtweg et al. 2021).

In their final form, semantic change test sets are simply lists of words where each word is accompanied either by a binary class label (where '1' means 'semantic shift' and '0' means 'no shift') or by a continuous value representing the degree of semantic change. The list is associated with two or more different time spans (e.g., the nineteenth and the twentieth century) and the corresponding corpora of texts published in the corresponding time spans used for creating the annotations. An automatic system is evaluated on its ability to predict the class label or the change score. As a rule, the classification predictions are evaluated using accuracy or an F-1 score, while the change scores are evaluated using the Spearman rank correlation between the predictions of the system and human annotations.

Unfortunately, manually annotated semantic change datasets (as DWUGs or in other forms) are still unavailable for the majority of the world's languages, and those that are available are rather small. Doubts have been expressed, for example, about whether one can trust Spearman rank correlations calculated on sets of 30 or 40 elements (Gonen

et al. 2020). Thus, the problem of evaluating approaches to semantic change modeling is far from being solved, and practitioners often rely on, or complement with, self-created test sets or manual analysis of the outcome of a system.

An interesting solution is the use of existing dictionaries or thesauri that indicate the year when a particular word sense was introduced. This approach was taken in the dataset presented in Cook et al. (2014) based on the *Macmillan English Dictionary for Advanced Learners* (MEDAL), and in the datasets introduced by Tahmasebi and Risse (2017a) and Tsakalidis et al. (2019) based mostly on the *Oxford English Dictionary*. In the same vein, van Aggelen et al. (2019) presented the large HiT dataset based on the *Historical Thesaurus of English*.[5] However, high-quality dictionaries (especially containing diachronic sense information) are still a scarce resource for the vast majority of the world's languages except English and lack grounding in any specific dataset. Grounding is necessary because the fact that a new sense has been documented does not necessarily mean that the sense can be detected in our available corpora.

Yet another evaluation strategy is to use the predicted diachronic semantic change to trace or predict real-world events like armed conflicts that took place in the corresponding time spans. Thus, event datasets (created and annotated by researchers in other fields of science such as history, political studies, and social studies) can serve as proxies for language change, contributing to the infrastructure of the field.

Finally, when lacking manually annotated test sets, one can turn to so-called 'synthetic evaluation', which is rooted in the field of word sense disambiguation (WSD), where artificially created 'ambiguous' pseudo-words have long been used to evaluate supervised algorithms (Schütze 1998). In WSD, pseudo-words are injected into real corpora to imitate synchronic lexical polysemy. In semantic change modeling, such pseudo-words are injected to imitate polysemy changing diachronically (e.g., a word gradually acquiring or losing a sense over time). Since these words are injected by a researcher and known by definition, the gold standard data emerges naturally. Synthetic evaluation was applied to semantic change detection by Dubossarsky et al. (2019) and Shoemark et al. (2019), among others. However, it should always be kept in mind that synthetic data follows a researcher's assumptions about how real semantic shifts should behave. It is never the same as real annotated data, and thus the conclusions drawn from synthetic evaluation should be taken with a grain of salt. This is why manually annotated diachronic word usage graphs (Schlechtweg et al. 2021) are still the most widely used resource to evaluate approaches to the detection of semantic changes.

To sum up, test datasets (benchmarks) annotated in a standard way for different languages and time periods form a substantial part of the research infrastructure in computational semantic change research.

2.4.3 Evaluation through shared tasks

The computational semantic change modeling landscape is, like many other modern NLP fields, heavily influenced by shared tasks, which are essentially research competitions to perform a task as well as possible. The goal is to evaluate a set of methods under the same conditions and on the same data. By using fair evaluation, we can learn which methods are more appropriate for different tasks and under different conditions, such as varying time periods, genres, or across multiple languages.

Organizers of a shared task publish the task description and a test dataset *without the correct answers* (sometimes a training dataset is also published). Participants then prepare their systems and use them to predict the answers for the published test set.

Each participating system is then ranked (automatically by the organizers) according to the accuracy of its predictions (forming 'leaderboards'). After the winner has been chosen and the shared task is closed, the correct answers are revealed. Shared tasks play an important role in establishing the state-of-the-art in natural language processing tasks (or, to put it simply, in finding the best systems). They often result in annotated datasets that help the field move forward long after the competition is over. They can also be used for training or fine-tuning purposes.

Semantic change modeling has seen several shared tasks in the past few years for English, German, Latin, Swedish (Schlechtweg et al. 2020), Italian (Basile et al. 2020), Russian (Kutuzov and Pivovarova 2021b), Spanish (Zamora-Reina, Bravo-Marquez, and Schlechtweg 2022), Finnish (Fedorova et al. 2024), and Chinese (Chen et al. 2023). Through these shared tasks, we can see the development of the field. In the earlier shared tasks, the leaderboards were dominated by static or 'type-based' word embeddings like word2vec (Mikolov et al. 2013). But starting from 2021, these are increasingly being replaced by contextualized or "token-based" approaches employing pretrained deep neural language models like BERT (Devlin et al. 2019), which are found to yield more accurate predictions (Montanelli and Periti 2023). These advances would be impossible without the established infrastructure of shared tasks for different languages. In addition, shared tasks lead to more attention being paid to languages other than English.

It should also be noted that those who participate in shared tasks are encouraged to publish their code and best models. This has led to a relative abundance of available diachronic models, ranging from static embeddings trained on specific decades of the Kubhist corpus (Hengchen and Tahmasebi 2021) and the COHA corpus (Kutuzov, Velldal, and Øvrelid 2022) to the contextualized Hist-BERT model pretrained on the same corpus (Wenjun Qiu and Xu 2022), and pre-trained multilingual models fine-tuned on Word-in-Context tasks (Cassotti et al. 2023).

There are other smaller-scale evaluation sets, such as the GEMS data set (Gulordava and Baroni 2011) and the Word Sense Change Test Set (Tahmasebi and Risse 2017b). The latter also aims to distinguish between different types of change and assigns a time period to each change. Neither of the above are, however, grounded in any corpus, which means that there is no guarantee that the change events will be found in any single diachronic corpus, nor at the time they are expected to be found.

At present, the shared tasks do not differentiate between change types other than gained or lost senses. Further work is needed to create shared tasks or evaluation data for the additional levels of change specified in Table 2, and for more genres of text. Importantly, evaluation of semantic change methods is not, and should not be, limited to testing on predetermined evaluation data, because a method that solves a specific task is not necessarily the most useful for, say, describing *what* has changed. Methods should therefore be tested in 'live' settings by applying them to different corpora and evaluating the outcome. More on evaluation can be found in Tahmasebi, Borin, and Jatowt (2021).

3 Open research questions

In the past decades, the study of computational semantic change or lexical semantic change detection (LSCD) has slowly emerged as an independent research field.

We have seen a large suite of models to represent meaning and detect semantic change. Although we have made a great deal of progress, there is still much to be done, both in detecting semantic change and in utilizing the results in research. In parallel with the development of better models (e.g., with better sense-differentiation or the ability to detect change over hundreds of time points), we need to test our models. Such testing should include large-scale semantic change discovery in lexicography and historical linguistic research, as well as the study of changing concepts in text-based research. We need best practices and good maps indicating which models and settings are needed for different kinds of data (both in terms of genre, time span, and research objectives), as well as evaluation data with a broader range of research objectives.

We also need models that are robust. Currently models depend on data processing routines (e.g., on whether the training or testing corpora were lowercased or not), exhibit substantial corpus bias, and have difficulties operating on levels more abstract than word or token similarities (e.g., at the level of lexicographic senses). Until these issues are resolved, the output of current semantic change detection models still needs human scrutiny, unless the downstream task at hand is tolerant of high levels of false positives.

Although the state of the field has improved significantly in recent years, much remains to be done in terms of evaluation. First, the overwhelming majority of publications still apply only to English data. The language coverage should be expanded to include more typologically and genetically diverse languages and more varied time spans. Second, there are nontrivial interactions between the statistical properties of gold scores and the performance of semantic change detection systems (Kutuzov and Giulianelli 2020). These interactions stand in the way of generalizing the results of the existing systems (as shown in shared tasks and leaderboards) to languages in general.

3.1 Detailed semantic change

The majority of the computationally aided studies on semantic change stop after having detected that change has occurred. However, there is a great need for more detailed analysis of the nature of the shift. This includes:

1. Sub-classification of types of semantic shifts (e.g., broadening, narrowing, metaphorization). This problem was to some degree addressed by Mitra et al. (2014), Tahmasebi and Risse (2017a), Giulianelli et al. (2023), and Cassotti, de Pascale, and Tahmasebi (2024), but much more work is required.
2. Identifying the source of a change (e.g., linguistic or extralinguistic causes). Detecting causation is closely linked to the division between linguistic drifts and cultural shifts, as explained in Hamilton, Leskovec, and Jurafsky (2016a). Again, manually annotated datasets of both linguistically motivated and culturally motivated semantic shifts are needed.
3. Identifying groups of words that change their meaning together in correlated ways. Some research in this direction was started by Dubossarsky, Weinshall, and Grossman (2016), who showed that verbs change more than nouns, and nouns change more than adjectives. Rodina et al. (2019) rejected the hypothesis that evaluative adjectives change faster than other adjectives. This is also naturally related to demonstrating the (non)existence of the 'laws of semantic change' and to studying the processes of co-lexification. Since the number of lexical groups is potentially infinite, the most important and interesting groups should be identified and analyzed.

3.2 Onomasiological change

Computational models of semantic change are tools to study diachronic trajectories of the correspondence between *form* and *meaning*. As we have seen in the previous sections, computational approaches are indeed equipped with ways to formally represent both these central objects of study. However, the focus of a study can be on either form or meaning.

Semasiological studies are anchored on form: they focus on the change, over time, of the meanings associated with the same linguistic expression. Most computational studies of semantic change address semasiological questions such as how the meaning of the term *controller* has changed from its first attested occurrence in the fifteenth century to today. This focus is largely due to the fact that distributional representations, the backbone of most approaches, lend themselves naturally to semasiological inquiry: given a word form and a set of usage examples collected from different time periods, it is straightforward to obtain and compare distributional representations.

Only a minority of studies are anchored, instead, on meaning – in other words, they answer onomasiological questions such as what you call a handheld peripheral device designed to provide input to a computer or a gaming console, and what this object was called in the 1960s. Onomasiological studies focus on changes in the set of linguistic expressions associated with a given concept across time periods. Computational modeling of onomasiological change requires creative solutions to invert the direction of the form–meaning mapping, which is the default in distributional semantics. Using a distributional semantic model is similar to looking up the meaning of a word in a dictionary. When using a dictionary, it has always been more difficult to find the word associated with a certain meaning of interest than to find the meaning of a target word. Perhaps for this reason, onomasiological change has been largely neglected in favor of the easier-to-model semasiological change, with the exception of work by Yao et al. (2018) and Tahmasebi et al. (2012), the latter with a focus on named entities.

Fortunately, while distributional semantic models are typically used to represent the meaning of specific forms, they also provide representations of 'anonymous' portions of the meaning space. One can therefore think of most computational approaches to onomasiological change as methods to retrieve the word forms that correspond, or are closest, to unnamed regions of interest in the distributional semantic space. This problem has been elegantly formulated as *temporal word analogy* (Szymanski 2017), *query reformulation* (Berberich et al. 2009), and *time-based synonyms* (Kanhabua and Nørvåg 2010). In more application-oriented communities, where onomasiological change is perceived as a challenge for information retrieval systems that operate with collections of documents spanning long time periods, it is also called *temporal counterpart search* (Iwai and Sumikawa 2017; Zhang et al. 2016). More work on onomasiological change is needed and will be an important complement to semasiological change.

3.3 Multi-modal semantic change detection

The lexical representations underlying the computational approaches presented so far are inferred from textual data alone. That is, the context of occurrence of a target word of interest is the few words surrounding it in a sentence, the whole sentence, a paragraph, or a document. They are thus grounded in the distributional hypothesis. With the advances made in generative language models, such as GPT-3 (Brown et al. 2020),

we are pushing the boundaries of the distributional hypothesis by extending the context to question-and-answer pairs. These pairs can, but do not need to, be found in close proximity in any text. This first major extension of the distributional hypothesis, together with the huge size of the pretrained data (500 billion words for GPT-3) and the almost equally huge set of parameters (175 billion for GPT-3) has great benefits for the modeling capacities of generative models.

Machine learning advances in fields such as computer vision and speech processing make it possible to take advantage of richer data records, which include information encoded through different modalities such as images and sound. Using multi-modal contexts of word usage, we can construct *grounded multi-modal semantic spaces*, as exemplified by GPT-4 which includes images in its representation space (OpenAI 2023). These additional multi-modal contexts can, for example, help disambiguate polysemous words and allow us to model abstract senses in a better way.

Still, text and images are not sufficient to fully capture the meaning of words or their context-specific meaning. Often, the complete concept is captured in the surrounding settings, physical or cultural. It might not be explicitly mentioned that women should not visit others unchaperoned, or that in some cultures and times dining alone is unusual and signals loneliness. Nor might it be possible to discern from the explicit communication what it means to say *Shhh* while nodding toward a sleeping child. Currently, we do not know what and how much can be captured in language models by grounding them in multi-modal content. Perhaps it is sufficient to discover that there are no mentions or no images of people dining alone to deduce that dining alone is unusual. But it is unlikely to ever capture how it might feel to be the one who dines alone in a world where others eye you suspiciously. Some of this information can be captured if and when we can access sensory data. When we can ground a word like *fear* with the emotional reaction that accompanies it in different situations, it is likely that we can get better representations, and thus more useful models.

SEE ALSO: Mechanisms of Change in Lexical Semantics; Onomasiological Variation: How We Name Extralinguistic Reality; Semasiological Variation; Syntagmatic and Paradigmatic Changes in Meaning.

Acknowledgments

This work has been partially funded by the project Towards Computational Lexical Semantic Change Detection supported by the Swedish Research Council (2019–2022; contract 2018-01184) and by the research program Change is Key! supported by Riksbankens Jubileumsfond (under reference number M21-0021).

Notes

1. The TIME corpus contains about 275,000 articles from TIME magazine from 1923 to 2006, https://www.english-corpora.org/time.
2. https://books.google.com/ngrams.
3. http://corpus.byu.edu/coha.
4. The CLARIN association maintains an (incomplete) list of historical corpora (mostly with long-term time spans) at https://www.clarin.eu/resource-families/historical-corpora.
5. https://ht.ac.uk.

References and Suggested Readings

Baayen, R. Harald, Fabian Tomaschek, Susanne Gahl, and Michael Ramscar. 2017. "The Ecclesiastes Principle in Language Change." In *The Changing English Language: Psycholinguistic Perspectives*, 21–48.

Banda, Juan M., Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, and Gerardo Chowell. 2020. "A Twitter Dataset of 150+ Million Tweets Related to Covid-19 for Open Research." Technical report, Panacea lab, April. This dataset will be updated bi-weekly at least with additional tweets, look at the GitHub repository for these updates.

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski. 2014. "Don't Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 1: 238–247.

Basile, Pierpaolo, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020. "DIACR-Ita @ EVALITA2020: Overview of the EVALITA2020 Diachronic Lexical Semantics (DIACR-Ita) Task." In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian: Final Workshop (EVALITA 2020) CEUR Workshop Proceedings (CEUR-WS.org)*.

Berberich, Klaus, Srikanta J. Bedathur, Mauro Sozio, and Gerhard Weikum. 2009. "Bridging the Terminology Gap in Web Archive Search." In *Twelfth International Workshop on the Web and Databases (WebDB 2009)*.

Betti, Arianna, and Hein Van den Berg. 2014. "Modelling the History of Ideas." *British Journal for the History of Philosophy* 22 (4): 812–835.

Blank, Andreas. 1997. *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen*. Berlin and Boston: Max Niemeyer Verlag. DOI: 10.1515/9783110931600.

Bloomfield, Leonard. 1933. *Language*. London: Allen & Unwin.

Bochkarev, Vladimir, Valery Solovyev, and Sören Wichmann. 2014. "Universals versus Historical Contingencies in Lexical Evolution." *Journal of The Royal Society Interface* 11 (101): 20140841.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. "Enriching Word Vectors with Subword Information." *Transactions of the Association of Computational Linguistics* 5: 135–146. http://aclweb.org/anthology/Q17-1010.

Bréal, Michel. 1899. *Essai de sémantique*, 2nd ed. Paris: Hachette.

Brochhagen, Thomas, Gemma Boleda, Eleonora Gualdoni, and Yang Xu. 2023. "From Language Development to Language Evolution: A Unified View of Human Lexical Creativity." *Science* 381 (6656): 431–436.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan et al. 2020. "Language Models Are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33: 1877–1901.

Brunila, Mikael, and Jack LaViolette. 2022. "What Company Do Words Keep? Revisiting the Distributional Semantics of J.R. Firth & Zellig Harris." In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4403–4417. DOI: 10.18653/v1/2022.naacl-main.327.

Bybee, Joan. 2006. *Frequency of Use and the Organization of Language*. Oxford: Oxford University Press.

Cassotti, Pierluigi, Stefano de Pascale, and Nina Tahmasebi. 2024. "Using Synchronic Definitions and Semantic Relations to Classify Semantic Change Types." In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, vol. 1: *Long Papers*. DOI: 10.48550/arXiv.2406.03452.

Cassotti, Pierluigi, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. "XL-LEXEME: WiC Pretrained Model for Cross-Lingual LEXical sEMantic changE." In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, vol. 2: *Short Papers*, 1577–1585. https://aclanthology.org/2023.acl-short.135.

Chen, Jing, Emmanuele Chersoni, Dominik Schlechtweg, Jelena Prokic, and Chu-Ren Huang. 2023. "ChiWUG: A Graph-Based Evaluation Dataset for Chinese Lexical Semantic Change Detection." In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, edited by Nina Tahmasebi, Syrielle Montariol, Haim Dubossarsky, Andrey Kutuzov, Simon Hengchen et al., 93–99. https://aclanthology.org/2023.lchange-1.10.

Choi, Hyunyoung, and Hal Varian. 2012. "Predicting the Present with Google Trends." *Economic Record* 88 (s1): 2–9.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek et al. 2020. "Unsupervised Cross-Lingual Representation Learning at Scale." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. DOI: 10.18653/v1/2020.acl-main.747.

Cook, Paul, Jey Han Lau, Michael Rundell, Diana McCarthy, and Timothy Baldwin. 2013. "A Lexicographic Appraisal of an Automatic Approach for Detecting New Word Senses." In *Proceedings of eLex* 2013, 49–65.

Cook, Paul, Jey Han Lau, Diana McCarthy, and Timothy Baldwin. 2014. "Novel Wordsense Identification." In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 1624–1635. https://aclanthology.org/C14-1154.

Corney, David, Dyaa Albakour, Miguel Martinez, and Samir Moussa. 2016. "What Do a Million News Articles Look Like? In *Proceedings of the First International Workshop on Recent Trends in News Information Retrieval Colocated with 38th European Conference on Information Retrieval (ECIR 2016), Padua, Italy, March 20*, 42–47. http://ceur-ws.org/Vol-1568/paper8.pdf.

Cuba Gyllensten, Amaru, Evangelia Gogoulou, Ariel Ekgren, and Magnus Sahlgren. 2020. "SenseCluster at SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection." In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 112–118. DOI: 10.18653/v1/2020.semeval-1.12.

Davies, Mark. 2012. "Expanding Horizons in Historical Linguistics with the 400-Million Word Corpus of Historical American English." *Corpora* 7 (2): 121–157.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. "Indexing by Latent Semantic Analysis." *Journal of the American Society for Information Science* 41 (6): 391.

Del Tredici, Marco, Raquel Fernández, and Gemma Boleda. 2019. "Short-Term Meaning Shift: A Distributional Exploration." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1: *Long and Short Papers*, 2069–2075. https://www.aclweb.org/anthology/N19-1210.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1: *Long and Short Papers*, 4171–4186. DOI: 10.18653/v1/N19-1423.

Dobrushina, Nina, and Michael Daniel. 2016. *Two Centuries in Twenty Words*. Moscow: NRU HSE (in Russian).

Dubossarsky, Haim, Yulia Tsvetkov, Chris Dyer, and Eitan Grossman. 2015. "A Bottom Up Approach to Category Mapping and Meaning Change." In *Proceedings of the NetWordS 2015 Word Knowledge and Word Usage*, 66–70.

Dubossarsky, Haim, Daphna Weinshall, and Eitan Grossman. 2016. "Verbs Change More Than Nouns: A Bottom-Up Computational Approach to Semantic Change." *Lingue e linguaggio* 15 (1): 7–28.

Dubossarsky, Haim, Daphna Weinshall, and Eitan Grossman. 2017. "Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models." In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1147–1156. http://aclweb.org/anthology/D17-1119.

Dubossarsky, Haim, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. "Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change."

In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, edited by Anna Korhonen, David Traum, and Lluís Màrquez, 457–470. DOI: 10.18653/v1/P19-1044.

Eger, Steffen, and Alexander Mehler. 2016. "On the Linearity of Semantic Change: Investigating Meaning Variation via Dynamic Graph Models." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, vol. 2: *Short Papers*, 52–58. DOI: 10.18653/v1/P16-2009.

Fedorova, Mariia, Timothee Mickus, Niko Tapio Partanen, Janine Siewert, Elena Spaziani, and Andrey Kutuzov. 2024. "AXOLOTL'24 Shared Task on Multilingual Explainable Semantic Change Modeling." In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change, Bangkok, Thailand*. Association for Computational Linguistics.

Finkelstein, Lev, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. "Placing Search in Context: The Concept Revisited." In *Proceedings of the 10th International Conference on World Wide Web*, 406–414.

Geeraerts, Dirk. 1997. *Diachronic Prototype Semantics: A Contribution to Historical Lexicology*. Oxford: Oxford University Press.

Giulianelli, Mario, Marco Del Tredici, and Raquel Fernández. 2020. "Analysing Lexical Semantic Change with Contextualised Word Representations." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3960–3973. DOI: 10.18653/v1/2020.acl-main.365.

Giulianelli, Mario, Andrey Kutuzov, and Lidia Pivovarova. 2022. "Do Not Fire the Linguist: Grammatical Profiles Help Language Models Detect Semantic Change." In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, 54–67. DOI: 10.18653/v1/2022.lchange-1.6.

Giulianelli, Mario, Iris Luden, Raquel Fernandez, and Andrey Kutuzov. 2023. "Interpretable Word Sense Representations via Definition Generation: The Case of Semantic Change Analysis." In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, vol. 1: *Long Papers*, 3130–3148. DOI: 10.18653/v1/2023.acl-long.176.

Gonen, Hila, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. 2020. "Simple, Interpretable and Stable Method for Detecting Words with Usage Change across Corpora." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 538–555. DOI: 10.18653/v1/2020.acl-main.51.

Gower, John C., and Garmt B. Dijksterhuis. 2004. *Procrustes Problems*. Oxford: Oxford University Press.

Grewal, Karan, and Yang Xu. 2021. "Chaining Algorithms and Historical Adjective Extension." *Computational Approaches to Semantic Change* 6: 189. DOI: 10.5281/zenodo.5040312.

Gulordava, Kristina, and 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. https://www.aclweb.org/anthology/W11-2508.

Hamilton, William, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. "Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 595–605. DOI: 10.18653/v1/D16-1057.

Hamilton, William, Jure Leskovec, and Dan Jurafsky. 2016a. "Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2116–2121. DOI: 10.18653/v1/D16-1229.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016b. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, vol. 1: *Long Papers*, 1489–1501. DOI: 10.18653/v1/P16-1141.

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

Hawkins, Robert D., Michael Franke, Michael C. Frank, Adele E. Goldberg, Kenny Smith, Thomas L. Griffiths, and Noah D. Goodman. 2023. "From Partners to Populations: A Hierarchical Bayesian Account of Coordination and Convention." *Psychological Review* 130 (4): 977.

Hengchen, Simon, and Nina Tahmasebi. 2021. "A Collection of Swedish Diachronic Word Embedding Models Trained on Historical Newspaper Data." *Journal of Open Humanities Data* 7: 1–7.

Hengchen, Simon, Ruben Ros, and Jani Marjanen. 2019. "A Data-Driven Approach to the Changing Vocabulary of the 'nation' in English, Dutch, Swedish and Finnish Newspapers, 1750–1950." DOI: 10.34894/AVBD7A, DataverseNL, V2.

Hill, Felix, Roi Reichart, and Anna Korhonen. 2015. "Simlex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation." *Computational Linguistics* 41 (4): 665–695. DOI: 10.1162/COLI_a_00237.

Hilpert, Martin, and Stefan Th. Gries. 2009. "Assessing Frequency Changes in Multistage Diachronic Corpora: Applications for Historical Corpus Linguistics and the Study of Language Acquisition." *Literary and Linguistic Computing* 24 (4): 385–401. DOI: 10.1093/llc/fqn012.

Homskiy, Daniil, and Nikolay Arefyev. 2022. "DeepMistake at LSCDiscovery: Can a Multilingual Word-in-Context Model Replace Human Annotators? In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, 173–179. DOI: 10.18653/v1/2022.lchange-1.18.

Hu, Renfen, Shen Li, and Shichen Liang. 2019. "Diachronic Sense Modeling with Deep Contextualized Word Embeddings: An Ecological View." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3899–3908. DOI: 10.18653/v1/P19-1379.

Iwai, Kazuhiro, and Yasunobu Sumikawa. 2017. "Detecting Counterpart Word Pairs across Time." In *Proceedings of the Second International Conference on Advanced Wireless Information, Data, and Communication Technologies*, 1–4.

Jatowt, Adam, and Kevin Duh. 2014. "A Framework for Analyzing Semantic Change of Words across Time." In *IEEE/ACM Joint Conference on Digital Libraries*, 229–238.

Juola, Patrick. 2003. "The Time Course of Language Change." *Computers and the Humanities* 37 (1): 77–96. http://www.jstor.org/stable/30204881.

Jurgens, David, and Keith Stevens. 2009. "Event Detection in Blogs Using Temporal Random Indexing." In *Proceedings of the Workshop on Events in Emerging Text Types*, 9–16. https://aclanthology.org/W09-4302.

Kanerva, Pentti, Jan Kristofersson, and Anders Holst. 2000. "Random Indexing of Text Samples for Latent Semantic Analysis." In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, vol. 22.

Kanhabua, Nattiya, and Kjetil Nørvåg. 2010. "Exploiting Time-Based Synonyms in Searching Document Archives." In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, 79–88.

Keidar, Daphna, Andreas Opedal, Zhijing Jin, and Mrinmaya Sachan. 2022. "Slangvolution: A Causal Analysis of Semantic Change and Frequency Dynamics in Slang." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, vol. 1 *Long Papers*, 1422–1442.

Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. "Temporal Analysis of Language through Neural Language Models." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 61.

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

Kutuzov, Andrey, and Mario Giulianelli. 2020. "UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic Change Detection." In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 126–134. DOI: 10.18653/v1/2020.semeval-1.14.

Kutuzov, Andrey, and Elizaveta Kuzmenko. 2016. "Cross-Lingual Trends Detection for Named Entities in News Texts with Dynamic Neural Embedding Models." In *First International Workshop on Recent Trends in News Information Retrieval Co-Located with 38th European Conference on Information Retrieval (ECIR 2016)*, 27–32.

Kutuzov, Andrey, and Lidia Pivovarova. 2021a. "RuShiftEval: A Shared Task on Semantic Shift Detection for Russian." In *Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference Dialogue*, 533–546. DOI: 10.28995/2075-7182-2021-20-XX-XX.

Kutuzov, Andrey, and Lidia Pivovarova. 2021b. "Three-Part Diachronic Semantic Change Dataset for Russian." In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, 7–13. DOI: 10. 18653/v1/2021.lchange-1.2.

Kutuzov, Andrey, Lidia Pivovarova, and Mario Giulianelli. 2021. "Grammatical Profiling for Semantic Change Detection." In *Proceedings of the 25th Conference on Computational Natural Language Learning*, 423–434. DOI: 10.18653/v1/2021. conll-1.33.

Kutuzov, Andrey, Samia Touileb, Petter Mæhlum, Tita Enstad, and Alexandra Wittemann. 2022. "NorDiaChange: Diachronic Semantic Change Dataset for Norwegian." In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2563–2572. https://aclanthology.org/2022.lrec-1.274.

Kutuzov, Andrey, Erik Velldal, and Lilja Øvrelid. 2017. "Tracing Armed Conflicts with Diachronic Word Embedding Models." In *Proceedings of the Events and Stories in the News Workshop*, 31–36. http://aclweb.org/anthology/W17-2705.

Kutuzov, Andrey, Erik Velldal, and Lilja Øvrelid. 2022. "Contextualized Embeddings for Semantic Change Detection: Lessons Learned." *Northern European Journal of Language Technology* 8. DOI: 10.3384/nejlt.2000-1533.2022.3478.

Lau, Jey Han, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. "Word Sense Induction for Novel Sense Detection." In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 591–601. https://aclanthology.org/E12-1060.

Leech, Geoffrey. 1992. "100 Million Words of English: The British National Corpus (BNC)." *Language Research* 28 (1): 1–13.

Liao, Xuanyi, and Guang Cheng. 2016. "Analysing the Semantic Change Based on Word Embedding." In *Natural Language Understanding and Intelligent Applications*, edited by Chin-Yew Lin, Nianwen Xue, Dongyan Zhao, Xuanjing Huang, and Yansong Feng, 213–223. Cham: Springer International.

Lijffijt, Jefrey, Tanja Säily, and Terttu Nevalainen. 2012. "CEECing the Baseline: Lexical Stability and Significant Change in a Historical Corpus." In *Studies in Variation, Contacts and Change in English*, vol. 10. Research Unit for Variation, Contacts and Change in English (VARIENG).

Martinc, Matej, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020. "Capturing Evolution in Word Usage: Just Add More Clusters?" In *Companion Proceedings of the Web Conference 2020*, 343–349.

Martinc, Matej, Petra Kralj Novak, and Senja Pollak. 2020. "Leveraging Contextual Embeddings for Detecting Diachronic Semantic Shift." In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4811–4819. https://aclanthology.org/2020.lrec-1.592.

McAuley, Julian John, and Jure Leskovec. 2013. "From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise through Online Reviews." In *Proceedings of the 22nd International Conference on World Wide Web*, 897–908. DOI: 10.1145/2488388.2488466.

McGillivray, Barbara, and Adam Kilgarriff. 2013. "Tools for Historical Corpus Research, and a Corpus of Latin." *New Methods in Historical Corpus Linguistics* 1 (3): 247–257.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray et al. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331 (6014): 176–182. DOI: 10.1126/science.1199644.

Mihalcea, Rada, and Vivi Nastase. 2012. "Word Epoch Disambiguation: Finding How Words Change over Time." In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, vol. 2: *Short Papers*, 259–263. www.aclweb.org/anthology/P12-2051.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. "Distributed Representations of Words and Phrases and Their Compositionality." *Advances in Neural Information Processing Systems* 26: 3111–3119.

Mitra, Sunny, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, and Pawan Goyal. 2014. "That's Sick Dude! Automatic Identification of Word Sense Change across Different Timescales." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, vol. 1: *Long Papers*, 1020–1029. DOI: 10.3115/v1/P14-1096.

Montanelli, Stefano, and Francesco Periti. 2023. "A Survey on Contextualised Semantic Shift Detection." *arXiv preprint arXiv:2304.01666*.

Montariol, Syrielle, Matej Martinc, and Lidia Pivovarova. 2021. "Scalable and Interpretable Semantic Change Detection." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4642–4652.

Noble, Bill, Asad Sayeed, Raquel Fernández, and Staffan Larsson. 2021. "Semantic Shift in Social Networks." In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, 26–37. DOI: 10.18653/v1/ 2021.starsem-1.3.

Nowak, Krzysztof. 2019. "*Tempus Mutatur*: Analysing Collocations of *tempus* 'time' with Distributional Semantic Models." *Words and Sounds* 1: 69–85.

Ohlsson, Claes, Victor Wåhlstrand Skärström, and Henrik Björck. 2022. "The Market as a Concept in Swedish Parliamentary Records from 1867 to 1970: A Mixed Methods Study." In *Digital Parliamentary Data in Action (DiPaDA 2022) Workshop, Uppsala, Sweden, March 15, 2022*, 22–34.

OpenAI. 2023. GPT-4 Technical Report. https://arxiv.org/abs/2303.08774.

Orlikowski, Matthias, Matthias Hartung, and Philipp Cimiano. 2018. "Learning Diachronic Analogies to Analyze Concept Change." In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 1–11.

Parker, Robert, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. "English Gigaword Fifth Edition LDC2011T07." Technical Report, Linguistic Data Consortium, Philadelphia.

Paul, Hermann. 1880. *Prinzipien der Sprachgeschichte*. Berlin: Walter de Gruyter.

Periti, Francesco, Pierluigi Cassotti, Haim Dubossarsky, and Nina Tahmasebi. 2024. "Analyzing Semantic Change through Lexical Replacements." In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, vol. 1: *Long Papers*, 1–21.

Periti, Francesco, and Stefano Montanelli. 2024. "Lexical Semantic Change through Large Language Models: A Survey." *ACM Computing Surveys* 56 (11): art. 282. DOI: 10.1145/3672393.

Periti, Francesco, and Nina Tahmasebi. 2024. "A Systematic Comparison of Contextualized Word Embeddings for Lexical Semantic Change." In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1: *Long Papers*, 4262–4282.

Perrone, Valerio, Marco Palma, Simon Hengchen, Alessandro Vatri, Jim Q. Smith, and 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. DOI: 10.18653/v1/W19-4707.

Perrone, Valerio, Simon Hengchen, Marco Palma, Alessandro Vatri, Jim Q. Smith, and Barbara McGillivray. 2021. "Lexical Semantic Change for Ancient Greek and Latin." *Computational Approaches to Semantic Change*, 287–310.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark et al. 2018. "Deep Contextualized Word Representations." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1: *Long Papers*, 2227–2237. DOI: 10.18653/v1/N18-1202.

Ramscar, Michael, Peter Hendrix, Cyrus Shaoul, Petar Milin, and Harald Baayen. 2014. "The Myth of Cognitive Decline: Non-Linear Dynamics of Lifelong Learning." *Topics in Cognitive Science* 6 (1): 5–42.

Rissanen, Matti. 1994. "The Helsinki Corpus of English Texts." In *Corpora across the Centuries*, 73–79. Leiden: Brill.

Rissanen, Matti, and Ossi Ihalainen. 1991. The Helsinki Corpus of English Texts. University of Helsinki, Department of English.

Rodina, Julia, Daria Bakshandaeva, Vadim Fomin, Andrey Kutuzov, Samia Touileb, and Erik Velldal. 2019. "Measuring Diachronic Evolution of Evaluative Adjectives with Word Embeddings: The Case for English, Norwegian, and Russian." In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 202–209. DOI: 10.18653/v1/W19-4725.

Ryskina, Maria, Ella Rabinovich, Taylor Berg-Kirkpatrick, David R. Mortensen, and Yulia Tsvetkov. 2020. "Where New Words Are Born: Distributional Semantic Analysis of Neologisms and Their Semantic Neighborhoods." *Proceedings of the Society for Computation in Linguistics* 3 (1): 43–52.

Sagi, Eyal, Stefan Kaufmann, and Brady Clark. 2009. "Semantic Density Analysis: Comparing Word Meaning across Time and Phonetic Space." In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, 104–111.

Sandhaus, Evan. 2008. "The *New York Times* Annotated Corpus Overview." *Linguistic Data Consortium, Philadelphia* 6 (12): e26752.

Schlechtweg, Dominik, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. "Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 2: *Short Papers*, 169–174. DOI: 10.18653/v1/N18-2027.

Schlechtweg, Dominik, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. "SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection." In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 1–23. DOI: 10.18653/v1/2020.semeval-1.1.

Schlechtweg, Dominik, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021. "DWUG: A Large Resource of Diachronic Word Usage Graphs in Four Languages." In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7079–7091. DOI: 10.18653/v1/2021.emnlp-main.567.

Schmid, Hans-Jörg. 2017. *Entrenchment, Memory and Automaticity: The Psychology of Linguistic Knowledge and Language Learning*. Washington, DC: De Gruyter Mouton.

Schütze, Hinrich. 1998. "Automatic Word Sense Discrimination." *Computational Linguistics* 24 (1): 97–123. https://aclanthology.org/J98-1004.

Shoemark, Philippa, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. 2019. "Room to Glo: A Systematic Comparison of Semantic Change Detection Approaches with Word Embeddings." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 66–76. DOI: 10.18653/v1/D19-1007.

Sommerauer, Pia, and Antske Fokkens. 2019. "Conceptual Change and Distributional Semantic Models: An Exploratory Study on Pitfalls and Possibilities." In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 223–233.

Stern, Gustaf. 1931. *Meaning and Change of Meaning; with Special Reference to the English Language*. Göteborg: Wettergren & Kerbers.

Szymanski, Terrence. 2017. "Temporal Word Analogies: Identifying Lexical Replacement with Diachronic Word Embeddings." In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, vol. 2: *Short Papers*, 448–453. DOI: 10.18653/v1/P17-2071.

Tahmasebi, Nina N. 2013. "Models and Algorithms for Automatic Detection of Language Evolution". PhD thesis, Gottfried Wilhelm Leibniz Universität Hannover. http://edok01.tib.uni-hannover.de/edoks/e01dh13/771705034.pdf .

Tahmasebi, Nina, Lars Borin, and Adam Jatowt. 2021. "Survey of Computational Approaches to Lexical Semantic Change Detection." 10.5281/zenodo.5040302.

Tahmasebi, Nina, Gerhard Gossen, Nattiya Kanhabua, Helge Holzmann, and Thomas Risse. 2012. "NEER: An Unsupervised Method for Named Entity Evolution Recognition." In *Proceedings of COLING* 2012, 2553–2568. https://aclanthology.org/C12-1156 .

Tahmasebi, Nina, and Thomas Risse. 2017a. "Finding Individual Word Sense Changes and Their Delay in Appearance." In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, 741–749. 10.26615/978-954-452-049-6_095.

Tahmasebi, Nina, and Thomas Risse. 2017b. "Word Sense Change Testset." *DOI*: 10.5281/zenodo.495572.

Traugott, Elizabeth, and Richard Dasher. 2001. *Regularity in Semantic Change*. Cambridge: Cambridge University Press.

Tripodi, Rocco, Massimo Warglien, Simon Levis Sullam, and Deborah Paci. 2019. "Tracing Antisemitic Language through Diachronic Embedding Projections: France 1789–1914." In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 115–125. DOI: 10.18653/v1/W19-4715.

Tsakalidis, Adam, Marya Bazzi, Mihai Cucuringu, Pierpaolo Basile, and Barbara McGillivray. 2019. "Mining the UK Web Archive for Semantic Change Detection." In *Proceedings of Recent Advances in Natural Language Processing Conference*, 1212–1221.

Ullmann, Stephen. 1962. *Semantics: An Introduction to the Science of Meaning*. Oxford: Blackwell.

van Aggelen, Astrid, Antske Fokkens, Laura Hollink, and Jacco van Ossenbruggen. 2019. "A Larger-Scale Evaluation Resource of Terms and Their Shift Direction for Diachronic Lexical Semantics." In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, 44–54. https://aclanthology.org/W19-6105.

Vatri, Alessandro, and Barbara McGillivray. 2018. "The Diorisis Ancient Greek Corpus: Linguistics and Literature." *Research Data Journal for the Humanities and Social Sciences* 3 (1): 55–65.

Vejdemo, Susanne. 2017. "Triangulating Perspectives on Lexical Replacement." PhD thesis, Department of Linguistics, Stockholm University. https://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-137874.

Vylomova, Ekaterina, Sean Murphy, and Nicholas Haslam. 2019. "Evaluation of Semantic Change of Harm-Related Concepts in Psychology." In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 29–34. DOI: 10.18653/v1/W19-4704.

Wenjun Qiu, and Yang Xu. 2022. "Histbert: A Pre-Trained Language Model for Diachronic Lexical Semantic Analysis." https://arxiv.org/abs/2202.03612v1.

Wijaya, Derry Tanti, and Reyyan Yeniterzi. 2011. "Understanding Semantic Change of Words over Centuries." In *Proceedings of the 2011 International Workshop on DETecting and Exploiting Cultural diversiTy on the Social Web*, 35–40.

Xu, Yang, and Charles Kemp. 2015. "A Computational Evaluation of Two Laws of Semantic Change." In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci 2015)*. https://lclab.berkeley.edu/papers/xu_kemp_2015_parallelchange.pdf.

Yao, Zijun, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. "Dynamic Word Embeddings for Evolving Semantic Discovery." In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 673–681. DOI: 10.1145/3159652.3159703.

Zalizniak, Anna. 2018. "The Catalogue of Semantic Shifts: 20 Years Later." *Russian Journal of Linguistics* 22 (4): 770–787.

Zamora-Reina, Frank D., Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. "LSCDiscovery: A Shared Task on Semantic Change Discovery and Detection in Spanish." In *Proceedings*

of the 3rd Workshop on Computational Approaches to Historical Language Change, 149–164. DOI: 10.18653/v1/2022.lchange-1.16.

Zhang, Yating, Adam Jatowt, Sourav Bhowmick, and Katsumi Tanaka. 2015. "*Omnia mutantur, nihil interit*: Connecting Past with Present by Finding Corresponding Terms across Time." In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, vol. 1: *Long Papers*, 645–655. DOI: 10.3115/v1/P15-1063.

Zhang, Yating, Adam Jatowt, Sourav Bhowmick, and Katsumi Tanaka. 2016. "The Past Is Not a Foreign Country: Detecting Semantically Similar Terms across Time." *IEEE Transactions on Knowledge and Data Engineering* 28 (10): 2793–2807. DOI: 10.1109/TKDE.2016.2591008.