Towards more complete solutions for Lexical Semantic Change: an extension to multiple time periods and diachronic word sense induction

Francesco Periti*
University of Milan
Via Celoria 18
20133 Milan, Italy
francesco.periti@unimi.it

Nina Tahmasebi*
University of Gothenburg
Renströmsgatan 6
40530 Göteborg, Sweden
nina.tahmasebi@gu.se

Abstract

Thus far, the research community has focused on a simplified computational modeling of semantic change between *two time periods*. This simplified view has served as a foundational block but is not a *complete* solution to the complex modeling of semantic change. Acknowledging the power of recent language models, we believe that now is the right time to extend the current modeling to *multiple time periods* and *diachronic word sense induction*. In this position paper, we outline several extensions of the current modeling and discuss issues related to the extensions.

1 Introduction

Lexical Semantic Change (LSC) is the problem of automatically identifying words that change their meaning over time (Periti and Montanelli, 2024; de Sá et al., 2024; Tahmasebi et al., 2021; Kutuzov et al., 2018; Tang, 2018). Conceptually, this problem implicitly involves a fundamental step of *diachronic word sense induction* to distinguish each individual sense of a word over all the *multiple time periods* of interest (Periti et al., 2023; Alsulaimani and Moreau, 2023; Alsulaimani et al., 2020; Emms and Jayapal, 2016; Tahmasebi, 2013). However, the computational challenges in handling large corpora and the absence of comprehensive benchmarks have in practice led to a simplified modeling focused on *two* time periods t_1 and t_2 only. These are either modeled individually t_1, t_2 or in a single time interval $\langle t_1, t_2 \rangle$ considering all the data jointly.

Typically, approaches over two time periods are assumed to be directly extendable to real scenarios involving multiple time periods. For example, approaches designed for a single interval $\langle t_1, t_2 \rangle$, can be iteratively re-executed across multiple, contiguous intervals $\langle t_1, t_2 \rangle$, $\langle t_2, t_3 \rangle$, …, $\langle t_{n-1}, t_n \rangle$ (Giulianelli et al., 2020). However, multiple re-executions presents a computational challenge that significantly escalates as the number of considered periods increases. Procedures that were initially considered optional steps to expedite modeling in two time periods become fundamental over multiple time periods. For instance, since words can occur thousands of times in a diachronic corpus, it becomes imperative to randomly sample a limited number of occurrences and to leverage hardware components, such as GPU processor units.

Due to the absence of diachronic lexicographic resources (e.g., dictionaries, thesauri), and the gap between a general resource and specific data, the modeling of word sense is commonly approached in an *unsupervised* manner. Clustering techniques are generally employed to aggregate usages of a specific word into clusters, with the idea that each cluster denotes a specific word meaning that can be recognized in the considered documents. However, clusters of usages (regardless of method of clustering) do not necessarily correspond to precise senses (Martinc et al., 2020), but typically represent noisy projections related to specific context (Periti and Montanelli, 2024). As a result, manual activity is always required to translate the automatically derived clusters into a *diachronic sense inventory*. This sense inventory is the basis for interpreting the identified semantic change and modeling sense evolution (see Figure 1). While automatic methods, such as keywords extraction (Kellert and Mahmud Uz Zaman, 2022), or generating definitions for word usages (Giulianelli et al., 2023), have been proposed to support cluster interpretation, a reliable interpretation still needs manual supervision. Therefore, when multiple time periods are considered, interpretability challenges increase several orders of magnitude, making the direct re-execution of existing approaches unsuitable for effectively detecting semantic change and the evolution of each individual word meaning (Periti et al., 2023, 2022).

*Authors contributed equally

Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change, pages 108–119

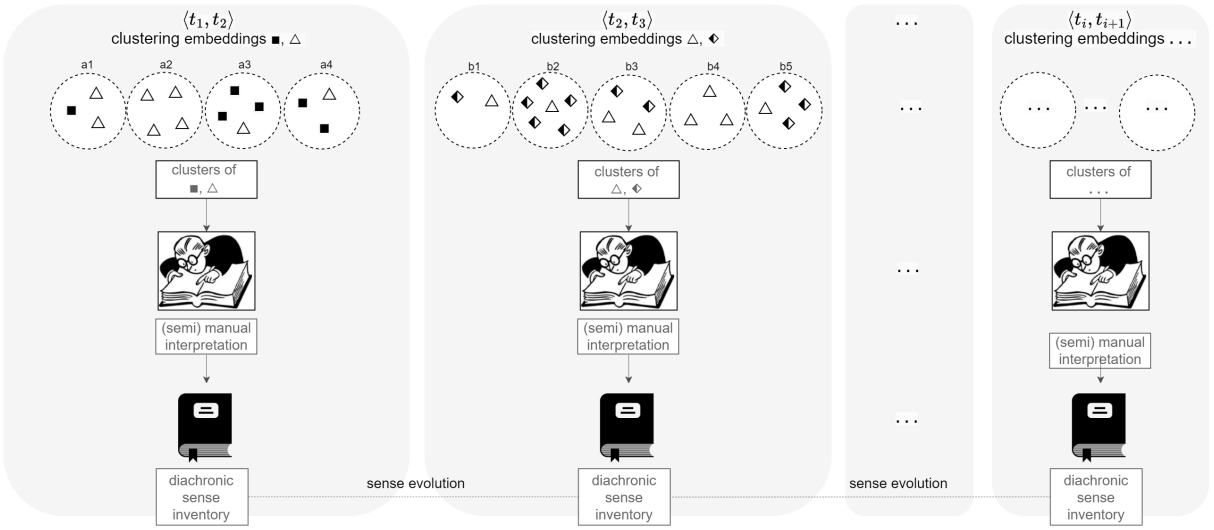

Figure 1: Word usages and their corresponding representations, for time period t_1, t_2, and t_3 are denoted with ■, △, ⬖, respectively. Typically, the clustering of representations is done for individual time interval (i.e., two time periods jointly) and manual supervision is required to translate the clusters of each time interval to a diachronic sense inventory. The amount of manual supervisions increase with the number of considered time intervals.

We thus argue that the *diachronic word sense induction* over *multiple time periods* inherent to LSC requires more careful considerations compared to the simplified modeling currently done. More efforts should be devoted to develop approaches for assisting text-based researchers like linguists, historians and lexicographers as much as possible.

Our original contribution

In this paper, we discuss the complexities inherent in modeling semantic change for each word sense individually over multiple time periods. We challenge the general assumption that conventional approaches designed to address LSC over two time periods are easily extendable over multiple time periods. Because currently, contextualized embeddings represents the preferred tool for addressing LSC (Periti and Tahmasebi, 2024), we will use these as an example. Our discussion is however more general, and can be applied regardless of which model is used to represent individual word usages – such as definitions (Giulianelli et al., 2023), co-occurrence vectors (Schütze, 1998), lexical replacements (Periti et al., 2024), or bag-of-substitutes (Kudisov and Arefyev, 2022) – or sense clusters in general as in Tahmasebi and Risse, 2017.

We advocate for an alternative modeling of LSC over multiple time periods, and specifically, we present i) five distinct approaches for *tracking* semantic change and the *evolution* of word meanings; and ii) three distinct settings for assessing semantic change over time. Our work has significant implications for both the computational modeling and the creation of benchmarks, contributing to the ongoing discussion presented by Periti and Montanelli (2024); Hengchen et al. (2021); Montariol et al. (2021) on the open challenges associated with modeling semantic change.

2 Background and related work

Since SemEval-2020 (Schlechtweg et al., 2020), there is an established evaluation framework for LSC to compare the performance of various models and approaches. However, given the substantial annotation efforts required to create reliable benchmarks over multiple time periods, the framework is typically adopted to create simplified benchmarks over two time periods, with gold labels for semantic change but without diachronic sense labels (Ling et al., 2023; Chen et al., 2023; Kutuzov et al., 2022a; Zamora-Reina et al., 2022; Kutuzov and Pivovarova, 2021; Basile et al., 2020; Schlechtweg et al., 2020).[1] In such benchmarks, the LSC problem is defined as follows.

2.1 Problem statement over two time periods

Given a diachronic corpus C containing a set of documents (e.g., sentences, paragraphs) from two time periods t_1 and t_2, the current modeling of LSC involves the following evaluation tasks:

[1] Kutuzov and Pivovarova, 2021 introduced a benchmark encompassing two time intervals. However, these intervals have been treated independently, leading to their consideration as two distinct sub-benchmarks over a single time interval.

i) to quantify the semantic change of words (i.e., *Graded Change Detection*);
ii) to recognize words that change their meaning by either gaining new ones or losing old ones (i.e., *Binary Change Detection, Sense Gain Detection, Sense Loss Detection*).

Words that change their meanings by means of gaining or loosing senses will have a high degree of (graded) semantic change, while words that have a high degree of graded change do not need to have lost or gained senses.

These tasks inherently involve the modeling of word meanings across t_1 and t_2. However, due to the lack of diachronic sense labels, researchers and practitioners tend to focus on addressing tasks **i)** and **ii)** without adequately tackling the challenges associated with modeling sense evolution.

2.2 State-of-the-art approaches to LSC

Thus far, computational approaches to solve the above tasks have followed a standard receipt using a *four-step pipeline* (Periti and Montanelli, 2024). Given a corpus C spanning two time periods t_1 and t_2, and a target word w:

1) extraction of the word occurrences from both t_1 and t_2;
2) computational representation of each occurrence (the current standard is to leverage pretrained contextualized embeddings);
3) word sense induction by aggregating embeddings with a clustering algorithm;
4) assessment of semantic change by leveraging a distance measure on the embeddings from t_1 and t_2.

Approaches are typically distinguished in *form-based* and *sense-based*. The former does not induce sense **(3)** but quantifies semantic change using **(1,2,4)**, either as a shift in the dominant meaning of w or in its degree of polysemy. There is thus no easy way to discern individual senses from the change score without integrating "close reading" by humans. Sense-based approaches remedies this by relying on all steps **(1-4)** but generally induce senses **(3)** in a *synchronic* way, without considering the temporal nature of the documents (Ma et al., 2024). That is, they consider all the documents from t_1 and t_2 available as a whole and perform a single clustering activity over the entire set of generated embeddings, regardless of their time origin.

2.3 Modeling senses through clusters

The clustering of representations via word sense induction, step **(3)** above, serves as a tool to operationalize word senses in an unsupervised fashion through unstructured text (Lake and Murphy, 2023). On one hand, this operationalization offers a flexible adaptation to the data under consideration and allows to derive senses that do not necessarily need to be aligned with available static lexicographic resources (Kilgarriff, 1997). For instance, senses derived from youth slang (Keidar et al., 2022), or scientific texts are unlikely to align with a general lexicon meant to cover the whole spectrum of a given language.

On the other hand, as computational models derive information from the contexts surrounding word tokens, sense modeling tends to emphasize word usages rather than word meanings (Tahmasebi and Dubossarsky, 2023; Kutuzov et al., 2022b). Thus, while ideally we would like each cluster to correspond to one, and only one sense, in practice, multiple clusters may correspond to different nuances of the same sense. This effect is further amplified when considering data from diverse time, domains, or genres, where distinct linguistic registers, styles, or co-occurrence patterns may results in different senses.

Additionally, the interpretation of clusters as senses requires a notion of (word) "meaning" that can both differ in the mind of humans according to social or cultural background and age, as well as in the varying usages of a word in context. Thus, the mapping of *clusters* to *senses* involves i) identifying commonalities on the usages of each cluster that may be judged differently, as well as ii) mapping these commonalities to word meanings. The outcome results in a *sense inventory*.

2.4 Modeling LSC over multiple time periods

Modeling LSC involves computationally deriving word senses progressively over time. This entails re-executing the steps **(1-4)** multiple times. At each execution i, a set of clusters is generated and humans are needed to identify and update the sense inventory. This involves mapping the clusters generated at the i-th execution to senses and aligning senses temporally.

The way senses align over time give us important insights into how word meanings change. Classifying *types* of semantic change has been long studied and different schema have been proposed (Blank,

1997; Bloomfield, 1933; Paul, 1880). Among others, common types of change include *broadening* of meaning (e.g., dog was used to refer to dogs of specific large and strong breeds), *narrowing* of meaning (e.g., `girl` was used to refer to people of either gender), *novel senses* (e.g., `rock` as a music genre) and *metaphorical* extensions (e.g., `surfing` the web). The result is a *diachronic* sense inventory with temporal information on the active senses at each time, as well as potential relationships between senses.

To facilitate the interpretation of semantic change and the evolution of word meaning, the current, *synchronic* modeling of senses can benefit from *diachronic* modeling encompassing both incremental word sense induction and cluster alignment (Kanjirangat et al., 2020). Aligning clusters computationally will allow the simultaneous interpretation of multiple clusters, thereby reducing the burden of manual supervision at each time period. Clusters aligned over time can potentially suggest the continuation of an active sense, as well as the broadening and narrowing of meanings. In contrast, clusters not aligned over time can reveal both the continuation of different senses, as well as types of semantic change, like metaphoric extension.

Thus far, word meanings have been modeled through conventional clustering algorithms such as Affinity Propagation (Martinc et al., 2020) or K-Means (Kobayashi et al., 2021). However, these algorithms were originally designed for one-time data clustering and are not inherently suited to handle temporal dynamics. Specifically, clusters generated at t_{i-1} can become mixed up when re-executing the algorithm with both previous data and new data points at time $\langle t_{i-1}, t_i \rangle$. Consequently, objects that were previously clustered together at time t_{i-1} may either remain in the same cluster or be reassigned to different clusters based on the updated data at time t_i. This dynamic nature complicates the task of tracking the history of specific clusters across different time periods, and can lead to the creation of noisy clusters, especially when new data points arrive according to a skewed distribution.

Diachronic sense clustering. Conventional unsupervised clustering algorithms do not incorporate the faithfulness properties typical in *incremental clustering* literature, where clustering activities at any given point in time should remain faithful to the already existing clusters as much as possible (Chakrabarti et al., 2006) while at the same time

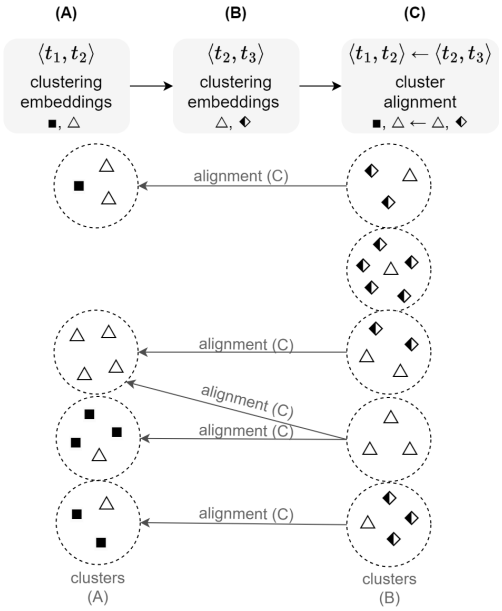

Figure 2: Clustering over consecutive time intervals.

be flexible to fit the new data. This would avoid dramatic change in clusters from one time-step to the next that do not derive from semantic change, but from differences in the underlying documents over time (Castano et al., 2024).

To this end, we argue that, for each target word, modeling LSC over time should involve *monitoring* the evolution of each individual senses across all the time periods under consideration, as well as *tracing* the types of each change. However, this extension is not straightforward; instead, it requires crucial time series analysis to mitigate potential noise introduced by the predictions of computational approaches (Kulkarni et al., 2015).

Monitoring and tracing word meaning evolution and semantic change require a careful consideration in the current *four-step pipeline* of sense-based approaches. As for scalability and interpretability issues related to **(1-3)**, suggestions and workaround are discussed in Periti and Montanelli, 2024; Montariol et al., 2021. In this paper, we further discuss the extension of steps **(3)** and **(4)** when considering multiple time points. In particular, we discuss *diachronic word sense induction* in Section 3, and *semantic change assessment* in Section 4.

3 Diachronic word sense induction

For the sake of simplicity, consider a diachronic corpus $\mathcal{C}$ spanning three general, consecutive time periods t_1, t_2, t_3, not necessarily contiguous. This simplification does not lead to any loss of information, but serves to aid the discussion in a clear

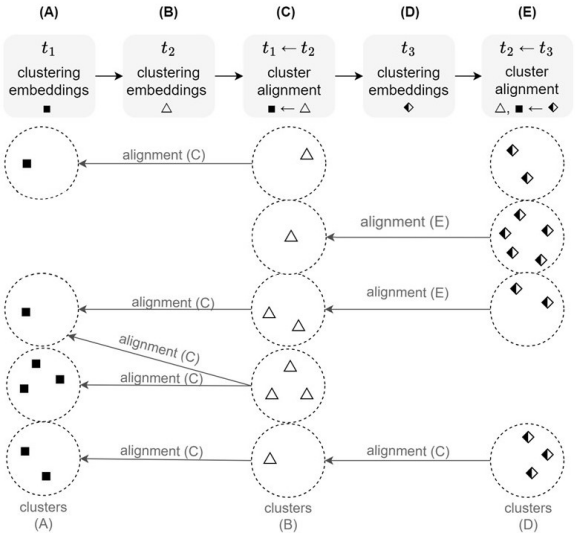

Figure 3: Clustering over consecutive time periods.

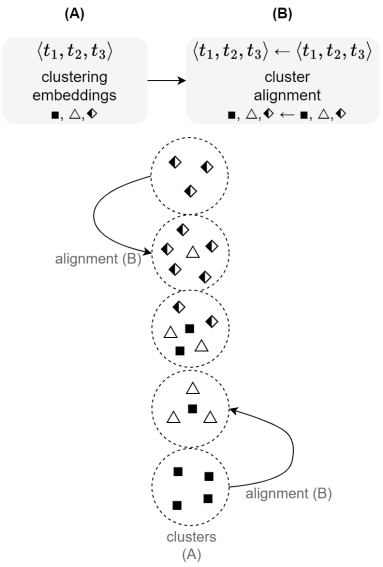

Figure 4: One-time clustering over all time periods.

and concise fashion. At the same time, three time points are easily extendable to the general case of tens or hundreds of time periods. Word usages, and their corresponding representations, for time period t_1, t_2, and t_3 are denoted with ■, △, ⬖, respectively. From here on, we will use contextualized embeddings as an example for contextualized representations. In the following, we present five different approaches for monitoring the evolution of word meanings and discuss suitability, and drawbacks.

3.1 Clustering over consecutive time intervals

Clustering algorithms used for *jointly* modeling senses over two time periods t_1 and t_2 can be progressively re-executed over consecutive pairs of time periods $\langle t_1, t_2 \rangle$ and $\langle t_2, t_3 \rangle$. To facilitate the interpretation of sense evolution, a cluster alignment step is thus required between consecutive re-executions. For instance, in Figure 2, the clusters generated in step (B) are linked to those generated in step (A) through a cluster alignment step (C) (Deng et al., 2019).

When clustering over consecutive time intervals $\langle t_1, t_2 \rangle, \ldots, \langle t_{n-1}, t_n \rangle$, the embeddings from $n-2$ time periods (all time periods but first and last) are clustered twice. For instance, consider the embeddings △ from t_2 in Figure 2: (A) they are first clustered with the embeddings ■ from t_1, and (B) then re-clustered with the embeddings ⬖ from t_3. When a limited number of word usages is available, this approach can potentially enhance the emergence of certain senses, as patterns of embeddings from t_{i-1} are reinforced by additional evidence (if present) from t_i. However, this compromises the faithfulness property, as embeddings from t_i can be clustered differently when considered jointly with t_{i-1} compared to when considered jointly with t_{i+1} (from a past and future perspective respectively).

3.2 Clustering over consecutive time periods

When a substantial number of documents is available for each time period, there is no need to cluster the embeddings of a time *interval* as a whole. Instead, the embeddings of each time *period* can be clustered individually, and a cluster alignment algorithm can be applied progressively to link the clusters across time periods (Kanjirangat et al., 2020; Montariol et al., 2021). This approach is represented in Figure 3. Step (A), (B), and (D) represents the application of a conventional clustering algorithm over the embeddings of time period t_1, t_2, t_3, respectively. Step (C) and (E) represents cluster alignment steps to link the clusters generated through step (B) to the cluster generated through step (A), and in turn, the clusters generated through step (D) to the cluster generated through step (B) (Deng et al., 2019).

Clustering over time periods involves a similar number of clustering activities and cluster alignment steps as clustering over time intervals. However, each clustering activity is more scalable, as it involves a smaller number of embeddings.

3.3 One-time clustering over all time periods

Embeddings from all the considered time periods can be clustered jointly in one single execution. For instance, in Figure 4 step (A), embeddings

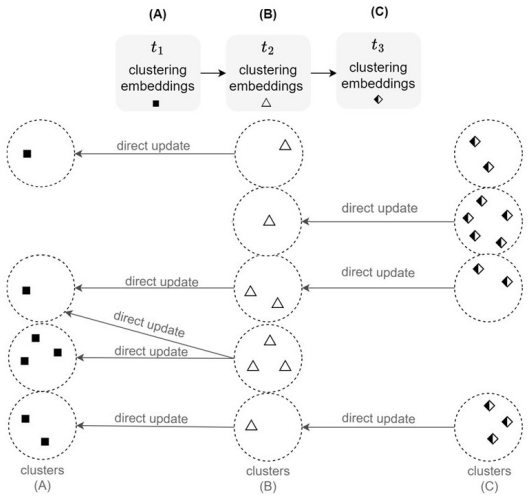

Figure 5: Incremental clustering over time periods.

■, △, ⬖ are clustered together as a whole. This single clustering activity results in clusters that may include embeddings from various combinations of time periods. For example, a cluster may include embeddings from a single, all, or selected time periods. A cluster alignment step (B) can be further executed to enable the modeling of sense evolution and change type.

When dealing with hundreds of time periods and a significant number of embeddings at once, clustering can be unfeasible due to scalability issues. In real scenarios, a diachronic corpus can be *dynamic* (Periti et al., 2022), where documents from subsequent time periods are not available as a whole but are progressively added (e.g., *posts* from social networks, Kellert and Mahmud Uz Zaman, 2022; Noble et al., 2021). In such scenarios this approach is thus not suitable as it would require re-execution of the clustering from scratch when new documents are added.

Furthermore, the use of conventional clustering algorithms is generally insensitive to the order of time periods, allowing embeddings of later time periods to influence pattern of the earlier time periods. This risks leading to a global view of word meaning while precluding a local view where smaller and gradual variation of individual senses as well as small sense clusters are missed. These issues can be mitigated by considering the temporal order of documents in the clustering activity (Smyth, 1996).

3.4 Incremental clustering over time periods

Incremental clustering algorithms are designed to effectively address the temporal nature of data (Kulkarni and Mulay, 2013). Thus, they are

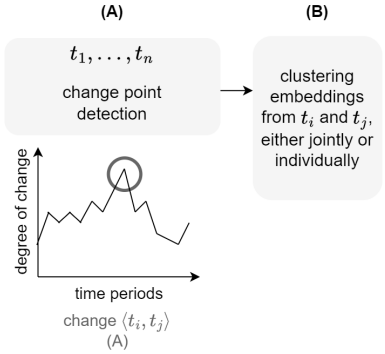

Figure 6: Scaling up with form-based approaches.

a suitable option to model the dynamic nature of language where temporal progression is key. When employed for diachronic word sense induction, they can efficiently and directly update the prior clustering results by processing and assimilating new data into existing clusters. The word usages observed in past time periods are consolidated into a set of clusters that constitute the *memory* of the word meanings observed thus far (Periti et al., 2022). This memory then serves as a foundation for understanding subsequent word usages in the current time period. Like Figure 4, Figure 5 represents similar steps (A-C) without alignment as clusters generated in step (A-C) are directly and consecutively updated.

Some of the incremental algorithms implement the faithfulness property in an *evolutionary* way: once a cluster has been created, it can only gain new members (i.e, word usages) but can never loose any members that have already been assigned to it. Meanwhile, the word usages observed in the present must be stratified or integrated over those from the past, that is, either be placed in existing clusters, or create new clusters. Other algorithms implement the faithfulness property in a more flexible way and enable small changes in past clusters when more evidence is available.

3.5 Scaling up with form-based approaches

Regardless of the complexity of each presented method, it is difficult to scale an approach to the level of whole vocabulary in a large corpus. In addition, some senses remain stable for a long time before they potentially change meaning, others never change. Therefore, clustering the senses during the stability periods of words is superfluous. To reduce computational needs and scale to the entire vocabulary, form-based approaches (without sense-induction) can be used to monitor stability allowing

the use of more powerful sense-based approaches only when there is indication of change.

By considering change only in the general usage of a word, form-based approaches reduce the semantic change problem significantly. Thus, they serve for two important purposes: first, they can be used to quantify the degree of change at the vocabulary level, and thus give us the opportunity to quantify change during different time periods (e.g., before and after WWI v. WWII); secondly, they can be used to find words and periods of interest.

Such a kind of stability monitoring can be done via change point detection (Kulkarni et al., 2015) and be integrated with diachronic sense modeling as shown in Figure 6. In particular, step A involves quantifying semantic change through form-based assessment to detect change points across the entire time span covered by the corpus. Step B involves modeling each individual sense of the word around the detected change point(s) through approaches presented in Section 3.1-3.4.

4 Semantic change assessment

The diachronic word sense induction is independent from the assessment of change at the level of senses or words. While the modeling of word meaning relies on the notion of word senses, the assessment of change depends on the research questions that we want investigate. E.g., considering a perfect sense inventory we may want to ask how many meanings have been lost and gained, and if change is more evident in some time intervals compared to others. The answer to these depend on the way we assess change.

Assessment of change, like sense induction, has focused on two time intervals which is the smallest unit over which we can quantify change. However, generalizing from two intervals to multiple intervals is not trivial and needs considerations that depend heavily on the kind of research question that is being asked, as well as the kind of data available. Short-term data contra long-term data, or small contra large data require different strategies for quantifying change. Here we present some possible strategies that extend to multiple time periods.

Assessment over consecutive time intervals represents a general way to assess semantic change over time $\langle t_1, t_2\rangle$, $\langle t_2, t_3\rangle$, …, $\langle t_{n-1}, t_n\rangle$. This kind of assessments can be affected by i) (random) fluctuations in the underlying corpus, where the coverage of topics can be heavily influenced by real-life events; and ii) noisy artifacts of the computational modeling, e.g., influenced by frequency. The use of time series analysis or statistical tests can reduce the effect of potential artifacts from the data and capture only significant changes evident in the time series (Liu et al., 2021; Kulkarni et al., 2015).

This assessment represents a useful solution for scenarios where the focus is on detecting immediate changes, such as in rapidly evolving fields or during specific events that might impact language usage. When comparing $\langle t_{i-1}, t_i\rangle$, the assumption is that all the active word meanings in t_i, except for the new or changed ones, are active also in t_{i-1}. However, some senses are periodic and an undesirable side-effect is that they may be detected as change each time they appear and disappear as they are not represented in t_{i-1}.

Pairwise assessment over time periods Sometimes research questions may be tailored to specific time intervals (e.g, *before* and *after* the time period t_i of the corona pandemic). Thus, this assessment aims to quantify the change across specific time intervals $\langle t_{i-1}, t_i\rangle$ and $\langle t_j, t_{j+1}\rangle$ such that $i < j$. This assessment is also useful for identifying changes in periodic senses when the periodicity of the sense is known. For example, the meaning of the term *gold* related to the Olympic games that take place every forth year.

This assessment is also useful when research questions are tailored to specific types of change irrespective of when the change occurs. For example, when a diachronic sense inventory is available, broadening or narrowing can be investigated regardless of their time-specific appearance.

When all possible time intervals are considered, this assessment is associated with a computational complexity of $\mathcal{O}(n^2)$ where n is the number of considered periods. However, it provides a broader view of how meaning evolves over different spans, capturing trends that may not be apparent in consecutive intervals. For example, gradual changes over time would not appear with assessment over consecutive time intervals as too little evidence would be present, but could appear as radical changes when larger gaps between intervals are used.

By considering all the possible time intervals it is also possible to quantify the **global** level of change over the whole corpus. This method is insensitive to the order of the time periods and is useful for capturing overarching trends and patterns in semantic change across the entire timeline.

Cumulative assessment over time When research questions focus on the novel senses gained at time period t_i, the comprehensive overview of active senses from the past must be considered $\bigcup_{j=1}^{i-1} t_j$. Instead of considering only consecutive or specific time intervals, each new time period should be compared with the full diachronic sense inventory. Cumulative assessment emphasizes the overall evolution of meaning, providing a holistic view of changes from the beginning to the end of the timeline. It is useful for consolidating the evidence across multiple time periods which would not suffice on their own. For example, when research questions focus on the novelty introduced in time period t_i compared to the past periods, the assessment of change should consider the cumulative evidence of the past as a single, large time period. Similar assessment can be employed when research questions want to compare a past time period t_i with respect to the following $\bigcup_{j=i+1}^{n-1} t_j$, until the n^{th} time period.

5 Discussion and conclusion

Computational modeling of semantic change has long been done in a simplified way due to the challenges related to modeling senses across multiple time periods. However, sense inventories and the type of change a word exhibits, are fundamental aspects for text-based researchers like historians, linguists and lexicographers, and therefore, the full complexity of semantic change must be taken into consideration in the computational modeling. Now that we have powerful language models like GPT-4 (OpenAI, 2023) and XL-LEXEME (Cassotti et al., 2023) there are no excuses for taking a simplistic view on the modeling of semantic change.

In this paper, we have presented possible extensions to expand on the simplistic view. These extensions have equal implications both for the computational modeling as for the generation of manually annotated benchmarks which has also been done over two time periods due to the sheer volume of required annotations.

Crucial for the usefulness of semantic change studies is a *diachronic sense inventory* where the different senses are linked together to capture semantic change type and linguistic relation. It is using the diachronic sense inventory that the majority of the research questions can be answered. These pertain both to linguistic, language-level questions, but also to societal and cultural enquiries where text can be used as evidence. How to best frame and store the diachronic sense inventory is still an open issue and requires involvement from the communities around computational modeling of semantic change, word sense induction and lexical semantics in general, as well as the text-based researchers that will use the outcome.

Human supervision is necessary to develop a reliable sense inventory. As diachronic corpora can span multiple time periods and contain millions of documents, automatic supervision support is mandatory to reduce manual efforts as much as is possible. In this regard, aligning similar clusters and detecting change types to speed up the interpretation process is as crucial as it is difficult. Employing different kinds of diachronic word sense induction and assessment as outlined here, will lead to different amounts of manual interaction.

Aligning clusters over time poses a very challenging task, as some clusters may represent outliers, time intervals may be characterized by different numbers of clusters, and multiple noisy (or nuanced) clusters denoting the same meaning may emerge. As a result, the cluster alignment often involves the discretization of a fuzzy problem (Kianmehr et al., 2010), that is the creation of new global clusters that encompass sets of fuzzy clusters. Furthermore, when cluster are aligned through a posteriori step rather than being linked and updated directly, the alignment process (worst case) involves comparing each cluster with every other cluster across all time periods. This risks amplifying the potential level of noise and require intricate decisions typically taken without any theoretical basis.

Thus far, the research community has focused more on the quantification of semantic change rather than the underlying word sense induction because form-based approaches consistently outperformed sense-based approaches. However, the clustering algorithms that have been employed do not take the temporal nature of documents into consideration, and we thus argue that they are not optimal for modeling word meaning over time.

In this paper, we have outlined several possible paths forward, both in terms of diachronic word sense induction and assessment of change. We have left methods for change type detection for future work. Each proposed path is suitable for different kinds of research questions and data. For example, by clustering embeddings over a whole corpus, smaller senses that would not appear in

sequential modeling can gain sufficient evidence in global clustering. Such a method is however computationally expensive. Other methods suffer from the problem that when only consecutive time periods are considered, slow and gradual shift risks being missed and over long time periods other strategies are more suitable. Among these methods, we strongly advocate for a shift towards incremental methods as these are currently the best fit to the LSC problem.

6 Limitations

This is a position paper and as such, we have not reported any experiments nor proposed concrete algorithms. Instead, we have outlined general weaknesses of the current methods in the field of computational modeling of semantic change and discussed possible ways forward. We believe that different kinds of solutions can be used for this purpose, spanning from different classes of clustering algorithms (e.g., evolutionary, Periti et al., 2023) to different classes of graphs and networks (e.g., temporal, Ma et al., 2024).

We have focused on unsupervised methods that induce senses through clustering of word representations. In particular, we have focused on contextualized representations, which represent the de facto standard, irrespective of the model that is used to generate the representations (e.g., Devlin et al., 2019; Hofmann et al., 2021; Cassotti et al., 2023). We only mention other methods such as word masking for lexical substitutions (Card, 2023; Arefyev and Zhikov, 2020) or previous paradigms such as the use of static embeddings (Shoemark et al., 2019). Typically, static embeddings, as well as methods based on SVD or PPMI (Hamilton et al., 2016), collapse all the meanings of a word into a single static vector, thus our proposals may not be considered suitable for such solutions even if dynamic word embeddings such as those presented by Bamler and Mandt (2017); Yao et al. (2018); Rudolph and Blei (2018) are used. However, we argue that the methods outlined in this paper are directly extendable to methods based on static embeddings where sense clusters are generated by looking at the top neighbors in the embedding space (Gonen et al., 2020).

We have not focused on how to detect the *type* of semantic change nor the *cause* of it, primarily due to space limitations. However, we believe that the methods outlined in this paper inherently offer ways to detect type, but not necessarily cause, of change. When we begin to target change types, we need evaluation benchmarks. Creating such benchmarks entail consolidating and digitizing the types of change offered in taxonomies as, for example, (Blank, 1997; Ullmann, 1957; Bloomfield, 1933; Stern, 1931; Bréal, 1904; Darmesteter, 1893; Paul, 1880; Reisig, 1839), such as the work started by Cassotti et al. (2024).

Acknowledgments

This work has in part been funded by the research program Change is Key! supported by Riksbankens Jubileumsfond (under reference number M21-0021).

References

Ashjan Alsulaimani and Erwan Moreau. 2023. Improving Diachronic Word Sense Induction with a Non-parametric Bayesian Method. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8908–8925, Toronto, Canada. Association for Computational Linguistics.

Ashjan Alsulaimani, Erwan Moreau, and Carl Vogel. 2020. An Evaluation Method for Diachronic Word Sense Induction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3171–3180, Online. Association for Computational Linguistics.

Nikolay Arefyev and Vasily Zhikov. 2020. BOS at SemEval-2020 Task 1: Word Sense Induction via Lexical Substitution for Lexical Semantic Change Detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 171–179, Barcelona (online). International Committee for Computational Linguistics.

Robert Bamler and Stephan Mandt. 2017. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 380–389. PMLR.

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

Andreas Blank. 1997. *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen*. Max Niemeyer Verlag, Berlin, Boston.

Leonard Bloomfield. 1933. *Language*. Holt, Rinehart and Winston, New York.

Michel Bréal. 1904. *Essai de Sémantique (Science des Significations)*. Hachette.

Dallas Card. 2023. Substitution-based Semantic Change Detection using Contextual Embeddings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 590–602, Toronto, Canada. Association for Computational Linguistics.

Pierluigi Cassotti, 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 (Volume 1: Long Papers)*, Bangkok, Thailand. Association for Computational Lingustics.

Pierluigi Cassotti, 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 (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.

Silvana Castano, Alfio Ferrara, Stefano Montanelli, and Francesco Periti. 2024. Incremental Affinity Propagation based on Cluster Consolidation and Stratification.

Deepayan Chakrabarti, Ravi Kumar, and Andrew Tomkins. 2006. Evolutionary Clustering. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 554–560, Philadelphia, PA, USA. Association for Computing Machinery.

Jing Chen, 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*, pages 93–99, Singapore. Association for Computational Linguistics.

Arséne Darmesteter. 1893. *La Vie des Mots Étudiée Dans Leurs Significations*. C. Delagrave.

Jader Martins Camboim de Sá, Marcos Da Silveira, and Cédric Pruski. 2024. Survey in characterization of semantic change.

Zhijie Deng, Yucen Luo, and Jun Zhu. 2019. Cluster Alignment With a Teacher for Unsupervised Domain Adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9943–9952.

Jacob Devlin, 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Martin Emms and Arun Kumar Jayapal. 2016. Dynamic Generative model for Diachronic Sense Emergence Detection. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1362–1373, Osaka, Japan. The COLING 2016 Organizing Committee.

Mario Giulianelli, 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*, pages 3960–3973, Online. Association for Computational Linguistics.

Mario Giulianelli, 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 (Volume 1: Long Papers)*, pages 3130–3148, Toronto, Canada. Association for Computational Linguistics.

Hila Gonen, 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*, pages 538–555, Online. Association for Computational Linguistics.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501.

Simon Hengchen, Nina Tahmasebi, Dominik Schlechtweg, and Haim Dubossarsky. 2021. *Challenges for Computational Lexical Semantic Change*, pages 341–372. Language Science Press, Berlin.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Dynamic contextualized word embeddings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6970–6984, Online. Association for Computational Linguistics.

Vani Kanjirangat, Sandra Mitrovic, Alessandro Antonucci, and Fabio Rinaldi. 2020. SST-BERT at SemEval-2020 Task 1: Semantic Shift Tracing by Clustering in BERT-based Embedding Spaces. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 214–221, Barcelona (online). International Committee for Computational Linguistics.

Daphna Keidar, 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 (Volume 1: Long Papers)*, pages 1422–1442, Dublin, Ireland. Association for Computational Linguistics.

Olga Kellert and Md Mahmud Uz Zaman. 2022. Using Neural Topic Models to Track Context Shifts of Words: a Case Study of COVID-related Terms Before and After the Lockdown in April 2020. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 131–139, Dublin, Ireland. Association for Computational Linguistics.

Keivan Kianmehr, Mohammed Alshalalfa, and Reda Alhajj. 2010. Fuzzy clustering-based discretization for gene expression classification. *Knowledge and Information Systems*, 24:441–465.

Adam Kilgarriff. 1997. "I Don't Believe in Word Senses". *Computers and the Humanities*, 31(2):91–113.

Kazuma Kobayashi, Taichi Aida, and Mamoru Komachi. 2021. Analyzing Semantic Changes in Japanese Words Using BERT. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 270–280, Shanghai, China. Association for Computational Lingustics.

Artem Kudisov and Nikolay Arefyev. 2022. BOS at LSCDiscovery: Lexical Substitution for Interpretable Lexical Semantic Change Detection. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 165–172, Dublin, Ireland. Association for Computational Linguistics.

Parag A. Kulkarni and Preeti Mulay. 2013. Evolve Systems using Incremental Clustering Approach. *Evolving Systems*, 4(2):71–85.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically Significant Detection of Linguistic Change. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 625–635, Florence, Italy. International World Wide Web Conferences Steering Committee.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic Word Embeddings and Semantic Shifts: a Survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Andrey Kutuzov and Lidia Pivovarova. 2021. Three-part Diachronic Semantic Change Dataset for Russian. In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 7–13, Online. Association for Computational Linguistics.

Andrey Kutuzov, Samia Touileb, Petter Mæhlum, Tita Enstad, and Alexandra Wittemann. 2022a. NorDiaChange: Diachronic Semantic Change Dataset for Norwegian. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2563–2572, Marseille, France. European Language Resources Association.

Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2022b. Contextualized Embeddings for Semantic Change Detection: Lessons Learned. In *Northern European Journal of Language Technology, Volume 8*, Copenhagen, Denmark. Northern European Association of Language Technology.

Brenden M Lake and Gregory L Murphy. 2023. Word Meaning in Minds and Machines. *Psychological Review*, 130(2):401–431.

Zhidong Ling, Taichi Aida, Teruaki Oka, and Mamoru Komachi. 2023. Construction of Evaluation Dataset for Japanese Lexical Semantic Change Detection. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 125–136, Hong Kong, China. Association for Computational Linguistics.

Yang Liu, Alan Medlar, and Dorota Glowacka. 2021. Statistically Significant Detection of Semantic Shifts using Contextual Word Embeddings. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 104–113, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xianghe Ma, Michael Strube, and Wei Zhao. 2024. Graph-based Clustering for Detecting Semantic Change Across Time and Languages.

Matej Martinc, 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*, WWW '20, page 343–349, Taipei, Taiwan. Association for Computing Machinery.

Syrielle Montariol, 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*, pages 4642–4652, Online. Association for Computational Linguistics.

Bill Noble, 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*, pages 26–37, Online. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 Technical Report.

Hermann Paul. 1880. *Prinzipien der Sprachgeschichte*. Niemeyer, Halle.

Francesco Periti, Pierluigi Cassotti, Haim Dubossarky, and Nina Tahmasebi. 2024. Analyzing Semantic Change through Lexical Replacements. In *Proceedings of the 62th Annual Meeting of the Association for Computational Linguistics*, Bangkok, Thailand. Association for Computational Linguistics.

Francesco Periti, Alfio Ferrara, Stefano Montanelli, and Martin Ruskov. 2022. What is Done is Done: an Incremental Approach to Semantic Shift Detection. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 33–43, Dublin, Ireland. Association for Computational Linguistics.

Francesco Periti and Stefano Montanelli. 2024. Lexical Semantic Change through Large Language Models: a Survey. *ACM Comput. Surv.* Just Accepted.

Francesco Periti, Sergio Picascia, Stefano Montanelli, Alfio Ferrara, and Nina Tahmasebi. 2023. Studying Word Meaning Evolution through Incremental Semantic Shift Detection: A Case Study of Italian Parliamentary Speeches.

Francesco Periti 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 (Volume 1: Long Papers)*, pages 4262–4282, Mexico City, Mexico. Association for Computational Linguistics.

Karl Christian Reisig. 1839. *Professor K. Reisig's Vorlesungen Über Lateinische Sprachwissenschaft*. Verlag der Lehnhold'schen Buchhandlung.

Maja Rudolph and David Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 1003–1011, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Dominik Schlechtweg, 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*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Hinrich Schütze. 1998. Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97–123.

Philippa Shoemark, 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)*, pages 66–76, Hong Kong, China. Association for Computational Linguistics.

Padhraic Smyth. 1996. Clustering sequences with hidden markov models. In *Advances in Neural Information Processing Systems*, volume 9. MIT Press.

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

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. *Survey of Computational Approaches to Lexical Semantic Change Detection*, pages 1–91. Language Science Press, Berlin.

Nina Tahmasebi and Haim Dubossarsky. 2023. Computational Modeling of Semantic Change.

Nina Tahmasebi and Thomas Risse. 2017. Finding Individual Word Sense Changes and their Delay in Appearance. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 741–749, Varna, Bulgaria. INCOMA Ltd.

Nina N. Tahmasebi. 2013. *Models and Algorithms for Automatic Detection of Language Evolution*. Ph.D. thesis, Gottfried Wilhelm Leibniz Universität Hannover.

Xuri Tang. 2018. A state-of-the-art of Semantic Change Computation. *Natural Language Engineering*, 24(5):649–676.

S. Ullmann. 1957. *The Principles of Semantics*. Glasgow University publications. Jackson.

Zijun Yao, 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*, WSDM '18, page 673–681, New York, NY, USA. Association for Computing Machinery.

Frank D. Zamora-Reina, 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*, pages 149–164, Dublin, Ireland. Association for Computational Linguistics.