Lexical Semantic Change through Large Language Models: a Survey

FRANCESCO PERITI, Computer Science, University of Milan, Milano, Italy
STEFANO MONTANELLI, Computer Science, University of Milan, Milano, Italy

Lexical Semantic Change (LSC) is the task of identifying, interpreting, and assessing the possible change over time in the meanings of a target word. Traditionally, LSC has been addressed by linguists and social scientists through manual and time-consuming analyses, which have thus been limited in terms of the volume, genres, and time-frame that can be considered. In recent years, computational approaches based on Natural Language Processing have gained increasing attention to automate LSC as much as possible. Significant advancements have been made by relying on Large Language Models (LLMs), which can handle the multiple usages of the words and better capture the related semantic change. In this article, we survey the approaches based on LLMs for LSC, and we propose a classification framework characterized by three dimensions: *meaning representation*, *time-awareness*, and *learning modality*. The framework is exploited to (i) review the measures for change assessment, (ii) compare the approaches on performance, and (iii) discuss the current issues in terms of scalability, interpretability, and robustness. Open challenges and future research directions about the use of LLMs for LSC are finally outlined.

CCS Concepts: • **Applied computing** → *Language translation*; • **Computing methodologies** → **Natural language processing**; **Lexical semantics;**

Additional Key Words and Phrases: Lexical semantics, lexical semantic change, semantic shift detection, large language models

ACM Reference Format:
Francesco Periti and Stefano Montanelli. 2024. Lexical Semantic Change through Large Language Models: a Survey. *ACM Comput. Surv.* 56, 11, Article 282 (June 2024), 38 pages. https://doi.org/10.1145/3672393

1 Introduction

In recent years, **Natural Language Processing (NLP)** has gained increasing attention due to the unprecedented capabilities of **Large Language Models (LLMs)** in facilitating linguistic analyses of human language. Among these analyses, the digitization of historical text corpora has recently prompted the use of LLMs to support and automate the study of language from a *diachronic* perspective. Language is viewed as a dynamic entity over time where words can undergo **lexical semantic change**—i.e., "*innovations which change the lexical meaning rather than the grammatical function of a form*" [15]. As an example, consider the change of the word `gay`, shifting from meaning `cheerful` to `homosexual` in the past century [53].

Authors' Contact Information: Francesco Periti (Corresponding author), Computer Science, University of Milan, Milano, Lombardia, Italy; e-mail: francesco.periti@unimi.it; Stefano Montanelli, Computer Science, University of Milan, Milano, Lombardia, Italy; e-mail: stefano.montanelli@unimi.it.

ACM 0360-0300/2024/06-ART282
https://doi.org/10.1145/3672393

This phenomenon has long been studied by linguists and other scholars in the humanities through time-consuming manual activities [13]. For instance, conventional methods for detecting, interpreting, and assessing semantic change primarily rely on "close reading" and require arranging hypotheses and testing procedures to build extensive catalogues of word descriptions. These analyses keep humans "in-the-loop" and have thus been narrowed in terms of the volume, genres, and time-frame that can be manually considered. A reliable computational approach that efficiently analyzes vast amounts of text with limited human intervention would be an extremely useful tool to assist researchers such as linguists, historians, and lexicographers. Such a tool would assist in creating and updating linguistic resources (e.g., lexicons, vocabularies, and thesauri) while also enhancing our understanding of historical and societal change reflected in language. For instance, consider the actual attention to topics like "politically correct": The word `retarded` has undergone semantic change over time, originally describing a neutral medical condition, but later acquiring offensive connotations when used as a derogatory insult [52, 101].

Modeling lexical semantic change through LLMs represents a new opportunity to scale up and automate the analysis as much as possible. Notably, distributional word representations (i.e., word embeddings) generated by LLMs emerged as an effective solution to capture the possible change over time in the meanings of a target word. Any embedding-based approach relies on the well-known distributional hypothesis in linguistics: "*You shall know a word by the company it keeps*" [39] and the foundational premise is that words (and word occurrences) that have similar meanings are encoded closely to each other in the embedding space [26, 44, 96].

The initial excitement for word embeddings prompted researchers and practitioners to model lexical semantic change by using *static* **Language Models (LMs)** [130]. These models have been widely adopted and the main approaches have been reviewed in three survey papers [74, 131, 132]. Typically, approaches based on static LMs encode a word into a single semantic embedding, which is then used to detect change in the dominant sense (i.e., word meaning) of the word, without considering its potential additional subordinate senses. However, subordinate senses can change on their own, regardless of their dominant sense. For example, considering the word `rock`, the `music` meaning evolved over time to encompass both `music` and a particular lifestyle, while the `stone` meaning remained unchanged [54]. Thus, the recent introduction of more advanced Transformer architectures [137] has established the use of LLMs as the preferred tool for modeling semantic change. In contrast with static LMs, approaches based on LLMs typically rely on different word representations according to the context in which a word occurs. For instance, different semantic vectors are generated when the word `rock` in the input sequence is used with the `music` connotation or with the `stone` meaning. This capability facilitates the modeling of linguistic *colexification* phenomena such as homonymy [124] and polysemy [42]. However, although more approaches based on LLMs are emerging, a classification framework and a corresponding survey of existing approaches are still missing.

In this article, we survey the main approaches based on LLMs to model the linguistic phenomenon of lexical semantic change through a corresponding NLP task called ***Lexical Semantic Change*** **(LSC)** (also known as *Semantic Shift Detection*). To this end, we propose a classification framework based on three dimensions of analysis, namely, *meaning representation*, *time-awareness*, and *learning modality*, that allows to effectively describe the featuring properties of both *form*- and *sense*-based approaches in which solutions are typically distinguished [45]. We also review assessment methods and metrics used by different approaches to measure and quantify the change of a word over a specific time interval. As a further contribution of our survey, the approaches to LSC are compared based on their performance against various reference benchmarks. This comparison aims to discuss the related performance issues and potential limitations.

The goal of our survey is to highlight the computational modeling of LSC and focus on LLMs, rather than the linguistic perspective and theory behind them.

The article is organized as follows: Section 2 presents the LSC problem with the related workflow and formalization. The proposed framework for approach classification is illustrated in Section 3. The classification of state-of-the-art approaches is discussed in Section 4. A comparative analysis of approach performance is provided in Section 5; issues related to scalability, interpretability, and robustness of the approaches are discussed in Section 6. In Section 7, we outline the open challenges and we give our concluding remarks.

2 Problem Statement

Consider a diachronic document corpus $C = \bigcup_{i=1}^{n} C_i$ where C_i denotes a set of documents (e.g., sentences, paragraphs) at time t_i; and a set of target words $\mathcal{W}$ occurring in the corpus C across the entire time span $[t_1, \ldots, t_n]$.

Modeling lexical semantic change typically involves:

— *word sense induction*: modeling the meaning(s) of each word $w \in \mathcal{W}$ in each time period $t_1, t_2, \ldots, t_n$;

— *semantic change detection*: identifying the words $w \in \mathcal{W}$ that change in meaning across all the contiguous time intervals, namely the pairs of time periods $\langle t_1, t_2 \rangle, \langle t_2, t_3 \rangle, \ldots, \langle t_{n-1}, t_n \rangle$.

For the sake of readability, in the following, we consider the LSC problem on a corpus $C = C_1 \cup C_2$ and the change assessment of a given target word $w \in \mathcal{W}$ on a single time interval $\langle t_1, t_2 \rangle$, from time period t_1 to time period t_2. This simplification enables to review the current state-of-the-art in a clear and concise fashion, while being easily extendable to the general case. As a matter of fact, the extension to the whole set of target words $\mathcal{W}$ as well as to all the contiguous time intervals can be obtained by re-executing a considered approach as many times as needed [45].

Different formulations of the problem are possibly depending on various research and assessment questions. The most popular are:

(1) **Graded Change Detection**: The goal is to quantify the extent to which a word w change in meaning between C_1 and C_2 [125].
(2) **Binary Change Detection**: The goal is to classify a word w as "stable" (without lost or gained senses) or "changed" (with lost or gained senses) between C_1 and C_2 [125].
(3) **Sense Gain Detection**: The goal is to recognize whether a word w gained meanings or not between C_1 and C_2 [143].
(4) **Sense Loss Detection**: The goal is to recognize whether a word w lost meanings or not between C_1 and C_2 [143].

2.1 The General Workflow

The approaches to LSC typically follow the four-step *workflow* presented in Figure 1. The initial **extraction** stage aims to select all the documents in the corpora containing occurrences (i.e., one or more) of the target word. We refer to these documents as *word usages*. The second **representation** stage has the goal to generate a semantic representation for each word occurrence. An optional **aggregation** stage can then be enforced to group multiple word representations into a single one for detecting similar usages and/or reducing the overall computational complexity. The final **assessment** stage consists in the application of a semantic measure to evaluate how the meanings of the word changed over time.

Word usage extraction. Consider the corpora C_1 and C_2 and the target word w. The goal of this stage is to extract all the contextual usages of w from C_1 and C_2. As the word meanings are

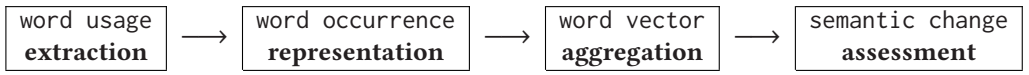

Fig. 1. A general workflow for modeling lexical semantic change through LLMs.

influenced by morphology and syntax [141], the extraction has to capture the occurrences of w in all its linguistic forms (e.g., singular/plural and gender forms, different verb tenses). For instance, a word may change in meaning only in one of its forms. An example is the Italian word `lucciola`, which was historically used with a euphemism for `prostitute`, a meaning that has now become obsolete. Nonetheless, the plural form `lucciole` has consistently retained the more stable sense of `fireflies` [77].

Word occurrence representation. The goal of this stage is to generate a word representation for each occurrence of the word w in C_1 and C_2. Ideally, the word representations of w should be similar for semantically similar word occurrences (i.e., usages) across different documents. An LLM is used to represent each word occurrence according to its context. Different types of representations can be used. Possible options are:

- **word embeddings**: a semantic vector in a multi-dimensional space that is directly generated by the Encoder of LLMs, such as BERT [33], RoBERTa [87], or ELMo [108].
- **lexical substitutes**: a bag of words that is generated by a Masked LLM such as BERT and RoBERTa to substitute a specific occurrence of w in a document [20]. These substitutes are supposed to replace a word without introducing grammatical errors or significantly changing its meaning. For example, suitable substitutes for the word `fly` in the sentence `a noisy fly sat on my shoulder` are `bug`, `beetle`, or `butterfly`; while suitable substitutes in the sentence `we will fly to London` are `walk`, `run`, or `bike` [70]. Alternatively, Causal LLMs such as GPT [18] and LLaMA [135] can be prompted to generate the substitutes [8, 103]. A **word embedding** vector for each occurrence of w can be computed over the substitutes (i.e., **bag-of-substitutes**) using measures like **Term Frequency-Inverse Document Frequency (TF-IDF)**.
- **sense definitions**: a descriptive interpretation that is generated by a Causal LLM to represent the occurrence of the word w in a particular document [47]. For example, an occurrence of the word `bank` may correspond to the definition of `a financial institution`, while another occurrence may correspond to `the edge of a river`. Alternatively, when available, lexical resources like WordNet [98] can be leveraged to obtain sense definitions. Sense definitions can be further processed by the Encoder of LLMs to generate less noisy **sense embedding** representations [69] or by **Natural Language Generation (NLG)** metrics such as BLEU, NIST, ROUGE-L, METEOR, or MoverScore [59].

Currently, **contextualized word embeddings** are the most widespread tool in LSC [79], with very few approaches using the other representations. Thus, we will use word embeddings as a reference for *word occurrence representation*. In the following, we denote the representation of the word w in the ith document of a corpus C_j as $e_{j,i}$, where $j \in 1, 2$. Then, the representation of the word w in a corpus C_j is defined as: $\Phi_j = \{e_{j,1}, \ldots, e_{j,z}\}$, with z being the cardinality of C_j, namely, the number of documents in C_j containing w. Finally, the sets of representation vectors generated for the word w at time t_1 and t_2 are denoted as Φ_1 and Φ_2, respectively.

Word vector aggregation. This stage is optionally executed and it has two main goals: (i) to recognize when different word occurrences convey a similar meaning and (ii) to reduce the number of elements to consider for change detection. To this end, clustering and averaging techniques are proposed for aggregating the generated word embeddings.

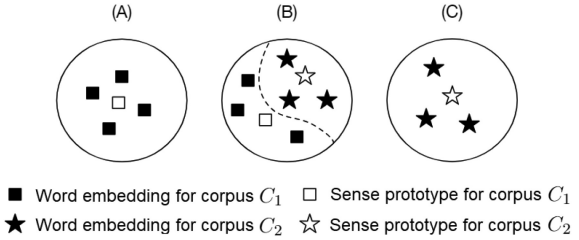

Fig. 2. Possible cluster composition for modeling word senses over time (from Reference [105]).

(i) **Clustering** techniques are employed to group similar word embeddings in a cluster, each one loosely denoting a specific word meaning. In some approaches, it is assumed that the corpus is *static*, meaning that all the documents in C_1 and C_2 are available as a whole. Then, a *joint* clustering operation is executed over the embeddings of $\Phi_1 \cup \Phi_2$ (e.g., Reference [92]). In other approaches, it is assumed that the corpus is *dynamic*, meaning that documents become available at different time periods and a *separate* clustering operation is performed over the embeddings of Φ_1 and Φ_2, individually (i.e., one exclusively on Φ_1 and another exclusively on Φ_2 embeddings). When a separate clustering is executed, the resulting clusters need to be aligned to recognize similar word meanings at different consecutive time periods (e.g., Reference [64]). To overcome the need for aligning clusters, an *incremental* clustering operation is employed to progressively group the embedding available at the different timesteps (e.g., Reference [106]). The result of clustering is a set of k clusters where the ith cluster is denoted as ϕ_i, and it can fall into one of the following cases (see Figure 2):
   – **(A)**: ϕ_i contains only embeddings from C_1;
   – **(B)**: ϕ_i contains a mixture of embeddings from both C_1 and C_2;
   – **(C)**: ϕ_i contains only embeddings from C_2.

   As a result, a cluster $\phi_i = \phi_{1,i} \cup \phi_{2,i}$ is composed by the union of two partitions $\phi_{1,i}$ and $\phi_{2,i}$ denoting the embeddings from Φ_1 and Φ_2, respectively. When a *joint* or *incremental* clustering is applied, the resulting clusters can belong to any of the above cases (i.e., A, B, and C). When a *separate* clustering is applied, the resulting clusters can just belong to A and C cases, meaning that $\phi_{2,i} = \emptyset$ and $\phi_{1,i} = \emptyset$, respectively.

(ii) **Averaging** techniques consist in determining a *prototypical* representation of the word w. As an option, a *word*-prototype can be computed by averaging all its embedding. In this case, *word*-prototypes μ_1 and μ_2 are created as the average embeddings of Φ_1 and Φ_2, respectively (e.g., Reference [73]). As an alternative option, averaging can be executed on top of the results of clustering. For each cluster, averaging is used to create a prototypical representation of all the cluster elements (i.e., the centroid of the cluster). In particular, *sense*-prototypes $c_{1,i}, c_{2,i}$ can be created for each cluster ϕ_i as the average embedding of its cluster partitions $\phi_{1,i}, \phi_{2,i}$, respectively (e.g., Reference [105]).

Semantic change assessment. This stage has the goal to measure the change on the meanings of the word w across the corpora C_1 and C_2 by considering the sets Φ_1 and Φ_2. In the literature, a number of functions are proposed for semantic change assessment. Distinctions can be made between measures that assess the change by considering the whole set of embedding representations Φ_i, by those that exploit the prototypical representations c_i and/or μ_i generated during the aggregation step through clustering and/or averaging. According to Reference [74], the definition of a rigorous, formal, mathematical model for representing the assessment functions used in LSC approaches is a challenging issue. In the following, we provide a formal definition of an abstract

function f with the goal of encompassing all existing assessment measures. The semantic change assessment $s = f(\cdot, \cdot, \cdot)$ is defined as follows:

$$f : \{\mathbb{R}^D\}^{(p_1+z_1\cdot\delta)}, \{\mathbb{R}^D\}^{(p_2+z_2\cdot\delta)}, \ c \rightarrow \mathcal{S},$$

where D is the dimension of the word vectors in Φ_1 and Φ_2; p_1, p_2 are the number of prototypical embeddings under consideration for C_1, C_2, respectively; z_1, z_2 are the number of vectors in Φ_1 and Φ_2, respectively; $\delta \in \{0, 1\}$ is a flag that allows to distinguish the approaches according to the kind of embedding used (i.e., original and/or prototypical); c is a counting function that determines the normalized number of embeddings in the cluster partitions $\phi_{1,i}$ and $\phi_{2,i}$, respectively.
The counting function c is defined as:

$$c : \{\mathbb{R}^D\}^{z_1}, \{\mathbb{R}^D\}^{z_2} \rightarrow \mathbb{R}^k, \mathbb{R}^k,$$

where k denotes the comprehensive number of k clusters obtained when a clustering operation is enforced during the aggregation stage. If a cluster ϕ_i contains embeddings only from Φ_1, then the corresponding count for C_2 will be equal to 0 and vice versa. When the clustering operation is not enforced, each embedding is mapped to a singleton group (i.e., $k = z_1 + z_2$).
The signature of f depends on the possible execution of an aggregation technique:

- *Clustering*. When the clustering operation is executed, then $p_1 = p_2 = 0$ and $\delta = 1$. This means that all the $z_1 + z_2$ embeddings in $\Phi_1 \cup \Phi_2$ are exploited for semantic change assessment (e.g., Reference [92]).
- *Averaging*. When the averaging operation is executed, then $p_1 = p_2 = 1$. In some approaches, $\delta = 0$ and this means that the function f is defined as a distance measure over prototypical representations (e.g., Reference [91]). In some other approaches, $\delta = 1$, and this means that f is defined as a distance measure over the original embeddings Φ and their prototypical representations (e.g., Reference [110]).
- *Clustering + Averaging*. When both clustering and averaging are performed, $p_1, p_2 > 0$ and δ can be both 0 or 1 as in the previous case (e.g., Reference [22]).

The output $\mathcal{S}$ is generally defined according to the formulation of the LSC problem.

- *Graded Change Detection*: $\mathcal{S} = \mathbb{R}$, with s quantifying the change of w between C_1 and C_2.
- *Binary Change Detection*: $\mathcal{S} = \{0, 1\}$, with s representing a binary score for "stable" (i.e., 0) and "changed" (i.e., 1), respectively.
- *Sense Gain Detection*: $\mathcal{S} = \{0, 1\}$, with s representing a binary score for not-gained (i.e., 0) and gained (i.e., 1), respectively.
- *Sense Loss Detection*: $\mathcal{S} = \{0, 1\}$, with s representing a binary score for not-lost (i.e., 0) and lost (i.e., 1), respectively.

Graded Change Detection is the most commonly considered formulation. Thus, in this survey, we focus on approaches that address LSC considering Graded Change Detection. It is worth noting that conceptually Binary Change Detection is not the binarization of Graded Change Detection. Indeed, even if a word does not gain/lose meanings (i.e., "stable" word), it can be associated with a high value of s due to other forms of semantic change, such as amelioration (change to positive connotation) and pejoration (change to negative connotation) [50]. However, in practice, Binary Change Detection is derived from Graded Change Detection by binarizing the graded s through a threshold θ (e.g., Reference [145]). We do not address Sense Gain and Sense Loss Detection, as they are relatively novel formulations.

For the sake of clarity, a summary of the notation used throughout this article is proved in Table 1.

Table 1. Summary of Notation Used in This Article

Notation	Definition
C	Diachronic document corpus
t_j	Time period jth
w	Target word
C_j	Set of documents at time t_j containing a word w
$\mathcal{W}$	Set of target words
$e_{j,i}$	Representation (i.e., embedding) of the word w in the ith document of a corpus C_j
Φ_j	Set of the representations of w in the corpus C_j
ϕ_i	ith cluster containing the representations of the word w
$\phi_{j,i}$	Subset of representations Φ_j in the cluster ϕ_i
μ_j	Prototypical representation of w for Φ_j
$c_{j,i}$	Prototypical representation of w for $\phi_{j,i}$

Table 2. A Classification Framework for Modeling Lexical Semantic Change

Meaning representation	Time-awareness	Learning modality
form-based	time-oblivious	supervised
sense-based	time-aware	unsupervised

3 A Classification Framework for LSC

A consolidated and widely accepted classification framework of approaches is not available. A basic framework is focused on the meaning representation of the words by distinguishing between *form*- and *sense*-based approaches [45, 112]. However, such a distinction is not universally recognized with a unique interpretation. Sometimes, these two categories are referred as *type*- and *token*-based, where averaging and clustering are enforced to aggregate embeddings, respectively [80, 125]. More recently, *average*- and *cluster*-based categories have been proposed to rename form and sense ones to highlight the method used for embedding aggregation [105].

In the following, this article proposes a comprehensive classification framework that extends the basic distinction between form- and sense-based approaches by introducing three dimensions of analysis, namely, `meaning representation`, `time-awareness`, and `learning modality` (see Table 2).

Meaning representation. Borrowing the distinction proposed by Reference [45], this dimension focuses on the meaning representation of a word. Two categories are defined:

— *form-based*: The meaning representation concerns the high-level properties of the target word w, such as its degree of polysemy or its dominant sense. When the polysemy is considered, the employed approaches do not enforce any aggregation stage, and the semantic change of w is assessed by measuring the degree of change on the embeddings Φ_1 and Φ_2 (i.e., change on the degree of polysemy). When the dominant sense is considered, all the meanings of w are collapsed into a single one on which the change is assessed. Typically, the embeddings Φ_1 and Φ_2 are averaged into corresponding word prototypes μ_1 and μ_2, respectively. In this case, the approaches focus on one meaning of w that can be considered as an approximation of the *dominant sense*, since, generally, it is the most frequent in the corpus, and thus the one most represented in the word prototype. We stress that form-based approaches are not able to represent how minor meanings *compete* and *cooperate* to change the dominant sense [58].

— *sense-based*: The meaning representation concerns the low-level properties of the target word w, such as its different context usages (i.e., its multiple meanings). All the senses of a word w are represented and considered in the change assessment, namely, the dominant sense and the minor ones. Typically, the embeddings Φ_1 and Φ_2 are aggregated into clusters, each one loosely representing a different meaning of w. Sense-based approaches allow to capture the changes over the different meanings of w as well as to interpret the word change (e.g., a new/existing meaning has gained/lost importance).

Time awareness. This dimension focuses on how the time information of the documents is considered by the employed LLM. Two categories are defined:

— *time-oblivious*: This category is based on the assumption that a document of time t adopts linguistic patterns that are known by the LLM and already characterize the language at the time t by its own. Thus, it is not needed that the LLM is aware of the time in which a document is inserted in the corpus. A time-oblivious approach is based on *the contextual nature of embeddings generated by the model, which by definition are dependent on the context that is always time-specific* [92].

— *time-aware*: This category is based on the assumption that the LLM is not capable of *adapting to time and generalizing temporally*, since they are *usually pre-trained on corpora derived from a snapshot of the web crawled at a specific moment in time* [116]. Thus, it is needed that the LLM is aware of the time in which a document is inserted in the corpus. As a result, a time-aware LLM encodes the time information as well as the linguistic context of a document while generating the word representations.

Learning modality. This dimension is about the possible use of external knowledge for describing and learning the word meanings to recognize. Two categories are defined:

— *supervised*: A form of supervision is enforced to inject external knowledge to support the change assessment. In addition to the text in the corpora C_1 and C_2, a lexicographic/manual supervision is employed. Lexicographic supervision refers to the use of dictionaries, vocabularies, or thesauri to support word sense induction and recognize the meaning of each word occurrence. This solution can be considered as an alternative to aggregation by clustering for meaning identification. Manual supervision involves using a human-annotated dataset (e.g., Word-in-Context dataset) with gold labels for training or fine-tuning the LLM [6].

— *unsupervised*: The change assessment is exclusively based on the text of the corpora C_1, C_2 without any external knowledge support. As a result, the word meanings to recognize emerge from the corpora and the change is completely assessed by exploiting unsupervised learning techniques. The use of aggregation by clustering is an example of unsupervised learning for meaning detection.

4 Approaches to LSC

In this section, the existing approaches in literature are reviewed according to the classification framework discussed in Section 3. In particular, the solutions are presented in Sections 4.1 and 4.2 according to the meaning representation of the considered target word, namely, *form-* and *sense-*based approaches, respectively. Moreover, Section 4.3 describes the so-called *ensemble* approaches, namely, approaches that are based on a combination of multiple form- and/or sense-based solutions.

For the sake of comparison, in each category (i.e., form, sense, ensemble), a summary table is provided to frame the literature papers according to the classification framework as well as to report additional descriptive features about the following aspects:

— *LLM*: the large language model used (e.g., BERT);
— *Training language*: the language of the dataset used for training the model. The possible options are *monolingual* to denote when training is executed on a single language or *multilingual* when more than one language is considered.
— *Type of training*: how the model is trained. Five categories are distinguished:
  – *trained*: The model is trained from scratch through a typical objective function(s);
  – *pre-trained*: The model has been pre-trained on a large dataset by other researchers, and it is directly used as an off-the-shelf solution instead of being trained from scratch;
  – fine-tuned for *domain-adaptation*: The model has been pre-trained on a large dataset by other researchers, then it is fine-tuned on new data through the same objective function;
  – fine-tuned for *incremental domain-adaptation*: The model is fine-tuned on the corpus of the first time period C_1. Then, it is re-tuned separately on the corpus C_2. The model at time t_2 is initialized with the weights from the model at time t_1, so both models are inherently related the one to the other;
  – *fine-tuned*: The model has been pre-trained on a large dataset by other researchers, then it is fine-tuned on new data through a different objective function.
— *Layer*: the architecture's layer(s) from which word representations are extracted;
— *Layer aggregation*: the type of aggregation used to synthesize the word representations extracted from different layers into a single embedding;
— *Clustering algorithm*: the clustering algorithm used in the aggregation stage;
— *Change function*: the function f used to detect/assess the semantic change;
— *Corpus language*: the natural language of the corpus in the considered experiments of change assessment (e.g., English, Italian, Spanish).

4.1 Form-based Approaches

According to Table 3, we note that most form-based approaches are time-oblivious. A few time-aware approaches have recently appeared, and they are all characterized by the adoption of a specific fine-tuning operation to inject time information into the model. All the current work leverages unsupervised learning modalities with the exception of Reference [6]. The aggregation stage is mostly based on averaging, while clustering is only enforced in Reference [12], where a cluster represents the dominant sense of the word w. In particular, in Reference [12], a word is considered as changing when clustering the embeddings Φ_1 and Φ_2 via K-means with $k = 2$ generates two groups where one of the two clusters contains at least 90% of the embeddings from one corpus only (i.e., C_1 or C_2).
In form-based approaches, the following change functions are proposed for measuring the semantic change s:

Cosine distance (CD). The change s is measured as the *cosine distance* (CD) between the word prototypes μ_1, μ_2 as follows:

$$CD(\mu_1, \mu_2) = 1 - CS(\mu_1, \mu_2), \tag{1}$$

where CS is the *cosine similarity* between the prototypes. Intuitively, the greater the $CD(\mu_1, \mu_2)$, the greater the change in the dominant sense of w.

Typically, the prototypes μ_1 and μ_2 are determined through aggregation by averaging over Φ_1 and Φ_2, respectively (e.g., Reference [91]). As a difference, in Reference [57], the prototype embedding μ_2 at timestep $t = 2$ is computed by updating the prototype embedding μ_1 at timestep $t = 1$ through a weighted running average (e.g., Reference [38]).

In Reference [91], the CD metric is employed in a multilingual experiment where the change is measured across a diachronic corpus with texts of different languages. This is the only example of cross-language change detection.

Table 3. Summary View of Form-based Approaches

Ref.	Time awareness	Learning modality	LLM	Training language	Type of training	Layer	Layer aggregation	Clustering algorithm	Change function	Corpus language
Arefyev et al. 2021	time-oblivious	supervised	XLM-R-large	multilingual	fine-tuned	last	–	–	APD	Russian
Beck 2020	time-oblivious	unsupervised	mBERT-base	multilingual	pre-trained	last two	average	K-means	CD	English, German, Latin, Swedish
Martinc et al. 2020a	time-oblivious	unsupervised	BERT-base, mBERT-base	monolingual, multilingual	domain-adaptation	last four	sum	–	CD	English, Slovenian
Horn 2021	time-oblivious	unsupervised	BERT-base, RoBERTa-base	monolingual	domain-adaptation, pre-trained	–	–	–	CD	English
Hofmann et al. 2021	time-aware	unsupervised	BERT-base	monolingual	fine-tuned	last	–	–	CD	English
Zhou and Li 2020	time-aware	unsupervised	BERT-base	monolingual	domain-adaptation	last four	sum	–	CD	English, German, Latin, Swedish
Rosin et al. 2022	time-aware	unsupervised	BERT-base, BERT-tiny	monolingual	fine-tuned	all, last, last four	average	–	CD, TD	English, Latin
Rosin and Radinsky 2022	time-aware	unsupervised	BERT-base, BERT-small, BERT-tiny	monolingual	fine-tuned	all, last, last four, last two	average	–	CD	English, German, Latin
Kutuzov and Giulianelli 2020	time-oblivious	unsupervised	BERT-base, ELMo, mBERT-base	monolingual, multilingual	domain-adaptation, incremental domain-adaptation, pre-trained, trained	all, last, last four	average	–	APD, CD, PRT	English, German, Latin, Swedish
Giulianelli et al. 2020	time-oblivious	unsupervised	BERT-base	monolingual	pre-trained	all	sum	–	APD	English
Keidar et al. 2022	time-oblivious	unsupervised	RoBERTa-base	monolingual	domain-adaptation	all, first, last	sum	–	APD	English
Pömsl and Lyapin 2020	time-aware	unsupervised	BERT-base, mBERT-base	monolingual, multilingual	fine-tuned	last	–	–	APD	English, German, Latin, Swedish
Kudisov and Arefyev 2022	time-oblivious	unsupervised	XLM-R-large	multilingual	pre-trained	–	–	–	APD	Spanish
Laicher et al. 2021	time-oblivious	unsupervised	BERT-base	monolingual	pre-trained	first, first + last, first four, last, last four	average	–	APD, APD-OLD/NEW, CD	English, German, Swedish
Wang et al. 2020	time-oblivious	unsupervised	mBERT-base	multilingual	pre-trained	last	–	–	APD, HD	Italian
Kutuzov 2020	time-oblivious	unsupervised	BERT-base, BERT-large, ELMo, mBERT-base	monolingual, multilingual	domain-adaptation, pre-trained	all, last, last four	average	–	APD, DIV, PRT	English, German, Latin, Swedish, Russian
Ryzhova et al.	time-oblivious	unsupervised	ELMo, **RuBERT** Kuratov and Arkhipov 2019	multilingual	pre-trained, trained	–	–	–	APD	Russian
Rodina et al. 2020	time-oblivious	unsupervised	ELMo, RuBERT	monolingual, multilingual	domain-adaptation	last	–	–	PRT	Russian
Liu et al. 2021	time-oblivious	unsupervised	BERT-base, **LatinBERT** Bamman and J. Burns 2020	multilingual, monolingual	domain-adaptation	last four	sum	–	CD	English, German, Latin, Swedish
Giulianelli et al. 2022	time-oblivious	unsupervised	XLM-R-base	multilingual	domain-adaptation	all	average	–	APD, PRT	English, German, Italian, Latin, Norwegian, Russian, Swedish
Laicher et al. 2020	time-oblivious	unsupervised	mBERT-base	multilingual	pre-trained	all, last four	average	–	APD	Italian
Qiu and Yang 2022	time-oblivious	unsupervised	BERT-base	monolingual	domain-adaptation pre-trained	last four	sum	–	CD	English
Periti et al. 2022	time-oblivious	unsupervised	BERT-base mBERT-base	monolingual, multilingual	pre-trained	last four	sum	–	CD, DIV	English, Latin
Montariol et al. 2021	time-oblivious	unsupervised	BERT-base mBERT-base	monolingual, multilingual	domain-adaptation	last four	sum	–	CD	English, German, Latin, Swedish

Missing information is denoted with a dash.

CD is also used in time-aware approaches. The integration of extra-linguistic information into word embeddings, such as time and social space, has been proposed in previous work based on static LMs [121, 144]. Recently, this integration has been also applied to contextualized embeddings [60, 119]. In Reference [56], a pre-trained LLM is fine-tuned to encapsulate time and social space in the generated embeddings. Then, the change s is assessed by computing the CD between embeddings generated by the original pre-trained model and the embeddings generated by the time-aware, fine-tuned model. In particular, in Reference [145], a *temporal referencing* mechanism is adopted to encode time-awareness into a pre-trained model. Temporal referencing is a pre-processing step of the documents that tags each occurrence of w in C_1 and C_2 with a special

marker denoting the corpus/time in which it appears [34, 37]. The embeddings of a tagged word are learned by fine-tuning the LLM for domain-adaptation. In this case, s is assessed by computing the CD between $\mu_{[1]}$ and $\mu_{[2]}$, where $[i]$ denotes w with the temporal marker t_i. Similarly to Reference [145], a time-aware approach is proposed in Reference [116] where a time marker is added to documents instead of words and the LLM is fine-tuned to predict the injected time information (i.e., time masking). This way, there is no need to add a tag for each target word and its various forms (e.g., singular, plural), thereby avoiding the inclusion of additional new tokens in the LLM's vocabulary. As an alternative, in Reference [117], a *temporal attention* mechanism is adopted to generate the embeddings Φ_1 and Φ_2 for calculating CD.

Inverted similarity over word prototype (PRT). This measure is proposed as an alternative to CD for improving the effectiveness of the change detection [73]. The *inverted similarity over word prototypes* (PRT) measure is defined as:

$$PRT(\mu_1, \mu_2) = \frac{1}{CS(\mu_1, \mu_2)}. \tag{2}$$

Time-diff (TD). This measure is designed for time-aware approaches, and it works on analyzing the change of polysemy of a word along time. It is based on the model capability to predict the time of a document, and it calculates the change s by considering the probability distribution of the predicted times [116]. Intuitively, a uniform distribution means that the association document-time is not strong enough to clearly entail a change. Instead, a non-uniform distribution means that there is evidence to predict the time of a document. Consider a document d, let $p_j(d)$ be the probability of d to belong to the time t_j. The function *time diff* (TD) is defined as the average difference of the predicted time probabilities:

$$TD(C_1, C_2) = \frac{1}{|C_1 \cup C_2|} \sum_{d_1 \in C_1, d_2 \in C_2} |p_1(d_1) - p_2(d_2)|. \tag{3}$$

The experiments conducted in Reference [116] demonstrate that TD outperforms CD in short-term semantic change when their performance is compared on the task of Graded Change Detection across various benchmarks. On the contrary, CD outperforms TD over long-term semantic change. Reference [116] argues that TD is less effective on long-term periods, since major differences in writing style emerge and the prediction of document-time associations is less reliable.

Average pairwise distance (APD). This measure exploits the variance of the contextualized representations Φ_1, Φ_2 to compute the semantic change assessment (i.e., variance on the word polysemy). As a difference with the previous measures, APD directly works on word embeddings without requiring any aggregation stage, namely, clustering nor averaging. The *average pairwise distance* (APD) is defined as follows:

$$APD(\Phi_1, \Phi_2) = \frac{1}{|\Phi_2||\Phi_2|} \cdot \sum_{e_{1,i} \in \Phi_1,\ e_{2,i} \in \Phi_2} d(e_{1,i}, e_{2,i}), \tag{4}$$

where d is an arbitrary distance measure (e.g., cosine distance, Euclidean distance, Canberra distance). According to the experiments performed in Reference [45], APD better performs when the Euclidean distance is employed as d. In Reference [67], APD is used over the embeddings Φ_1 and Φ_2 by applying a dimensionality reduction through the **Principal Component Analysis (PCA)**. In Reference [67], experiments on both slang and non-slang words are performed through causal analysis to study how distributional factors (e.g., polysemy, frequency shift) influence the change s. The results show that slang words experience fewer semantic change than non-slang words.

In Reference [70], lexical substitutes are used to assess s. A set of lexical substitutes is generated by leveraging a masked LLM (e.g., XLM-R) and word representations Φ_1, and Φ_2 are computed as *bag-of-substitutes*. Then, APD is finally computed over Φ_1, and Φ_2 to assess s.

APD is also used in a time-aware approach described in Reference [110], where a pre-trained BERT model is fine-tuned to predict the time period of a sentence. APD is finally used to measure the change between the embeddings extracted from the fine-tuned LLM.

In Reference [6], APD is employed to measure the change s over the embeddings Φ_1 and Φ_2 extracted from a supervised **Word-in-Context model (WiC)** [109]. This LLM is trained to reproduce the behavior of human annotators when they are asked to evaluate the similarity of the meaning of a word w in a pair of given sentences from C_1 and C_2, respectively. The embeddings Φ_1 and Φ_2 are extracted from the trained WiC model for calculating the final APD measure.

Average of average inner distances (APD-OLD/NEW). The APD-OLD/NEW measure is presented in Reference [81] as an extension of APD, and it estimates the change s as the average degree of polysemy of w in the corpora C_1 and C_2, respectively. The *average of average inner distances* (APD-OLD/NEW) is defined as:

$$APD\text{-}OLD/NEW(\Phi_1, \Phi_2) = \frac{AID(\Phi_1) + AID(\Phi_2)}{2}, \tag{5}$$

where AID is the *average inner distance,* and it measures the degree of polysemy of w in a specific time frame by relying on the APD measure, namely, $AID(\Phi_1) = APD(\Phi_1, \Phi_1)$ and $AID(\Phi_2) = APD(\Phi_2, \Phi_2)$, respectively.

Hausdorff distance (HD). The change s is measured as the *Hausdorff distance* (HD) between the word embeddings Φ_1 and Φ_2. Similarly to APD, HD directly works on word embeddings without requiring any aggregation stage. HD relies on the Euclidean distance d to measure the difference between the embeddings of w in C_1 and C_2, and it returns the greatest of all the distances d from one embedding $e_1 \in \Phi_1$ to the closest embedding $e_2 \in \Phi_2$ or vice versa. The HD measure is defined as follows:

$$HD(\Phi_1, \Phi_2) = \max\left(\sup_{e_1 \in \Phi_1} \inf_{e_2 \in \Phi_2} d(e_1, e_2), \sup_{e_2 \in \Phi_2} \inf_{e_1 \in \Phi_1} d(e_2, e_1) \right). \tag{6}$$

The experiments performed in Reference [138] show that HD is sensitive to outliers, since it is based on infimum and supremum, thus an outlier embedding may largely affect the final s value.

Difference between token embedding diversities (DIV). Similar to APD, this measure assesses the change s by exploiting the variance of the contextualized representation Φ_1 and Φ_2. As a difference with APD, the *difference between token embedding diversities* (DIV) leverages a coefficient of variation calculated as the average of the cosine distances d between the embeddings Φ_1 and Φ_2 and their prototypical embeddings μ_1 and μ_2, respectively [72]. The intuition is that when w is used in just one sense, its embeddings tend to be close to each other, yielding a low coefficient of variation. On the opposite, when w is used many different senses, its embeddings are distant to each other, yielding to a high coefficient of variation. DIV is defined as the absolute difference between the coefficient of variation in C_1 and C_2:

$$DIV(\Phi_1, \Phi_2) = \left| \frac{\sum_{e_1 \in \Phi_1} d(e_1, \mu_1)}{|\Phi_1|} - \frac{\sum_{e_2 \in \Phi_2} d(e_2, \mu_2)}{|\Phi_2|} \right|. \tag{7}$$

In Reference [72], the experiments show that when the coefficient of variation is low, the prototypical embeddings μ_1 and μ_2 successfully represent the meanings of the given word w. On the opposite, when the coefficient of variation is high, the prototypical embeddings μ_1 and μ_2 do not provide a relevant representation of the w meanings.

Table 4. Summary View of Sense-based Approaches

Ref.	Time awareness	Learning modality	LLM	Training language	Type of training	Layer	Layer aggregation	Clustering algorithm	Change function	Corpus language
Hu et al. 2019	time-oblivious	supervised	BERT-base	monolingual	pre-trained	last	–	–	MNS	English
Rachinskiy and Arefyev 2021	time-oblivious	supervised	XLM-R-base	multilingual	fine-tuned, pre-trained	–	–	–	APD	Russian
Rachinskiy and Arefyev 2022	time-oblivious	supervised	XLM-R-base	multilingual	fine-tuned, pre-trained	last	–	–	APD, JSD	Spanish
Periti et al. 2022	time-oblivious	unsupervised	BERT-base, mBERT-base	monolingual, multilingual	pre-trained	last four	sum	AP, APP, IAPNA	JSD, PDIS, PDIV	English, Latin
Montariol et al. 2021	time-oblivious	unsupervised	BERT-base, mBERT-base	monolingual, multilingual	domain-adaptation	last four	sum	K-means, AP	JSD, WD	English, German, Latin, Swedish
Rodina et al. 2020	time-oblivious	unsupervised	mBERT-base, ELMo	monolingual, multilingual	domain-adaptation	last	–	K-means, AP	JSD MS	Russian
Kanjirangat et al. 2020	time-oblivious	unsupervised	mBERT-base	multilingual	pre-trained	last four	concatenation	K-means	CSC, JSD	English, German, Latin, Swedish
Giulianelli et al. 2020	time-oblivious	unsupervised	BERT-base	monolingual	pre-trained	all	sum	K-means	ED, JSD	English
Arefyev and Zhikov 2020	time-oblivious	unsupervised	XLM-R-base	multilingual	domain-adaptation	–	–	AGG	CDCD	English, German, Latin, Swedish
Kashleva et al. 2022	time-oblivious	unsupervised	BERT-base	monolingual	domain-adaptation	all	sum	K-means	APDP	Spanish
Martinc et al. 2020c	time-oblivious	unsupervised	BERT-base, mBERT-base	monolingual, multilingual	domain-adaptation	last four	sum	K-means, AP	JSD	English, German, Latin, Swedish
Kutuzov and Giulianelli 2020	time-oblivious	unsupervised	BERT-base, ELMo, mBERT-base	monolingual, multilingual	domain-adaptation, incremental domain-adaptation, pre-trained	all, last, last four	average	AP	JSD	English, German, Latin, Swedish
Giulianelli et al. 2022	time-oblivious	unsupervised	XLM-R-base	multilingual	domain-adaptation	all	average	AP	JSD	English, German, Italian, Latin, Norwegian, Russian, Swedish
Wang et al. 2020	time-oblivious	unsupervised	mBERT-base	multilingual	domain-adaptation	last	–	GMMs, K-means	JSD	Italian
Keidar et al. 2022	time-oblivious	unsupervised	RoBERTa-base	monolingual	domain-adaptation	all, first, last	sum	AP, K-means, GMMs	ED, JSD	English
Karnysheva and Schwarz 2020	time-oblivious	unsupervised	ELMo, mELMo	monolingual, multilingual	pre-trained	all	–	K-means, DBSCAN	JSD	English, German, Latin, Swedish
Cuba Gyllensten et al. 2020	time-oblivious	unsupervised	XLM-R-base	multilingual	pre-trained	last	–	K-means	JSD	English, German, Latin, Swedish
Rother et al. 2020	time-oblivious	unsupervised	mBERT-base, XLM-R-base	multilingual	pre-tuned	last	–	BIRCH, DBSCAN, GMMs, HDBSCAN	JSD	English, German, Latin, Swedish

Missing information is denoted with a dash.

4.2 Sense-based Approaches

According to Table 4, we note that all the sense-based approaches are time-oblivious and that fine-tuning is sometimes adopted, but mainly for domain-adaptation purposes. Most papers leverage unsupervised learning modalities. Only a few exceptions employ a lexicographic supervision (i.e., References [58, 113, 114]). As a difference with form-based, sense-based approaches usually enforce clustering in the aggregation stage. The aggregation by averaging is only exploited in References [58, 100, 105], where sense prototypes are computed on top of the results of a clustering operation.

When clustering is adopted, the function f that calculates the change s can be directly defined over the embeddings Φ_1 and Φ_2. As an alternative, the function f can be defined over the distribution of the embeddings in the resulting clusters (i.e., *cluster distribution*). In this case, as a result of the clustering operation, a counting function c is used to determine two cluster distributions p_1 and p_2 that represent the normalized number of embeddings in the cluster partitions $\phi_{1,i}$ and $\phi_{2,i}$, respectively (see Section 2). The ith value $p_{j,i}$ in p_j (with $j \in \{1, 2\}$) represents the number of

embeddings of $\phi_{j,i}$ in the ith cluster, namely: $p_{j,i} = \frac{|\phi_{j,i}|}{|\Phi_j|}$. Finally, the function f is defined as a compound function $f = g \circ c$, where the result of the c function is exploited by a change function g that works on the cluster distributions p_1 and p_2.
In sense-based approaches, the following change functions are proposed for measuring the semantic change s:

Maximum novelty score (MNS). This measure exploits the cluster distributions p_1 and p_2 by leveraging the idea that the higher is the ratio between the number of embeddings Φ_1 and Φ_2 in a cluster, the higher is the semantic change of the considered word w. The *maximum novelty score* (MNS) is defined as:

$$MNS(p_1, p_2) = \max\{NS(p_{1,1}, p_{2,1}), \ldots, NS(p_{1,k}, p_{2,k})\}, \tag{8}$$

where $NS(p_{1,i}, p_{2,i}) = p_{1,i}/p_{2,i}$ is the *novelty score* proposed in Reference [28], and k is the number of clusters produced as a result of the aggregation stage.

In Reference [58], MNS is employed as a change measure in a supervised learning approach. In particular, a lexicographic supervision (i.e., the Oxford English dictionary) is employed to provide the meanings of the target word w. Each word occurrence in Φ_1 and Φ_2 is associated with the closest meaning of the dictionary according to the cosine distance. As a result, for each word/dictionary meaning, a cluster of word embeddings is defined and MNS is exploited to calculate the overall change.

Maximum square (MS). This measure is an alternative to MNS to assess the change of s. The intuition of MS is that slight changes in cluster distributions p_1 and p_2 may occur due to noise and do not represent a real semantic change [115]. The *maximum square* (MS) aims at identifying strong changes in the cluster distributions. As a difference with MNS, the square difference between $p_{1,i}$ and $p_{2,i}$ is used to capture the degree of change instead of the **novelty score (NS)**:

$$MS(p_1, p_2) = \max_i \left(p_{1,i} - p_{2,i}\right)^2 \tag{9}$$

Jensen-Shannon divergence (JSD). This measure extends the **Kullback-Leibler (KL)** divergence, which calculates how one probability distribution is different from another. The *Jensen-Shannon divergence* (JSD) calculates the change s as the symmetrical KL score of the cluster distributions p_1 from p_2, namely:

$$JSD(p_1, p_2) = \frac{1}{2}\left(KL(p_1||M) + KL(p_2||M)\right), \tag{10}$$

where KL is the Kullback-Leibler divergence and $M = (p_1 + p_2)/2$.

JSD is also used in approaches where aggregation by clustering is performed separately over the embeddings Φ_1 and Φ_2 [64]. As a result, the clusters need to be aligned to determine the distributions p_1 and p_2 before the JSD calculation. As a difference with Reference [64], an evolutionary clustering algorithm is employed in Reference [105] to apply the JSD measure without requiring any alignment step over the resulting clusters.

As a final remark, JSD can be employed to measure the change s over more than two time periods. However, the experiments in Reference [45] show that the JSD effectiveness over a single time period outperforms the version over more time periods, since JSD is insensitive to the order of the temporal intervals.

Coefficient of semantic change (CSC). This measure is proposed as an alternative to JSD, where the difference over the weighted number of elements in $\phi_{1,i}$ and $\phi_{2,i}$ for each cluster i is employed to replace KL in measuring the change [64]. The *coefficient of semantic change* (CSC)

is defined as follows:

$$CSC(p_1, p_2) = \frac{1}{P_1 \cdot P_2} \sum_{k=1}^{K} |P_2 \cdot p_{1,k} - P_1 \cdot p_{2,k}|, \tag{11}$$

where $P_j = \sum_{i=1}^{k} p_{j,i}$ is the weight of each cluster distribution and k is the number of clusters.

Cosine distance between cluster distributions (CDCD). As a further alternative of JSD, this measure assesses the change s by considering the cluster distributions p_1 and p_2 as vectors and by applying the cosine distance over them to assess the semantic change s. The *cosine distance between cluster distributions* (CDCD) is defined as follows:

$$CDCD(p_1, p_2) = 1 - \frac{p_1 \cdot p_2}{\|p_1\| \times \|p_2\|}. \tag{12}$$

In Reference [7], CDCD is calculated between the cluster distributions p_1 and p_2 obtained by enforcing clustering over bag-of-substitutes (see the description of Reference [7] in Section 4.1).

Entropy difference (ED). This measure is based on the idea that the higher is the uncertainty in the interpretation of a word occurrence due to the w polysemy in C_1 and C_2, the higher is the semantic change s. The intuition is that high values of ED are associated with the broadening of a word's interpretation, while negative values indicate a narrowing interpretation [45]. The *entropy difference* (ED) is defined as follows:

$$ED(p_1, p_2) = \eta(p_1) - \eta(p_2), \tag{13}$$

where $\eta(p_j)$ is the degree of polysemy of w in the corpus C_j, which is calculated as the normalized entropy of its cluster distribution p_j:

$$\eta(p_j) = \log_K \left(\prod_{k=1}^{K} {p_{j,i}}^{-p_{j,i}} \right).$$

As shown in Reference [45], ED is not capable of properly assessing s when new usage types of w emerge, while old ones become obsolescent at the same time, since it may lead to no entropy reduction.

Cosine distance between semantic prototypes (PDIS). This measure is presented in Reference [105] as an extension of the CD measure adopted by form-oriented approaches. The idea of PDIS is that the aggregation by averaging over cluster prototypes can be employed to produce summary descriptions of the cluster contents (i.e., *semantic prototypes*). The *cosine distance between semantic prototypes* (PDIS) is defined as the CD between $\bar{c}_1$, $\bar{c}_2$, that is:

$$PDIS(\bar{c}_1, \bar{c}_2) = 1 - \frac{\bar{c}_1 \cdot \bar{c}_2}{\|\bar{c}_1\| \times \|\bar{c}_2\|}, \tag{14}$$

where $\bar{c}_1$ and $\bar{c}_2$ are semantic prototypes defined as the average embeddings of all the sense prototypes $c_{1,i}$ and $c_{2,i}$, respectively.

Difference between prototype embedding diversities (PDIV). This measure is presented in Reference [105] as an extension of the DIV measure adopted by form-oriented approaches. PDIV leverages the same intuition of PDIS, namely, the semantic prototypes can be employed to calculate the coefficient of ambiguity of w by measuring the difference between a semantic prototype $\bar{c}_j$ and each sense prototype $c_{j,i}$. The *difference between prototype embedding diversities* (PDIV) is defined as the absolute difference between these ambiguity coefficients:

$$PDIV(\Psi_1, \Psi_2) = \left| \frac{\sum_{c_{1,k} \in \Psi_1} d(c_{1,k}, \bar{c}_1)}{|\Psi_1|} - \frac{\sum_{c_{2,k} \in \Psi_2} d(c_{2,k}, \bar{c}_2)}{|\Psi_2|} \right|, \tag{15}$$

where Ψ_1 and Ψ_2 denote the set of sense prototypes of $c_{1,i}$ and $c_{2,i}$, respectively.

Average pairwise distance (APD). In addition to form-based approaches (see Section 4.1), the APD measure is exploited to assess s also in sense-based approaches. In References [113, 114], APD is applied to the contextualized embeddings Φ_1 and Φ_2 extracted from a fine-tuned XLM-R model. In particular, an English corpus is used to fine-tune the pre-trained LLM to select the most appropriate WordNet's definition for each word occurrence [14]. As a result of the fine-tuning, both WordNet's definitions and word occurrences are embedded in the same vector space, and the meaning of any word occurrence can be induced by selecting the closest definition in the vector space. In Reference [113], the zero-shot, cross-lingual transferability property of XLM-R is exploited to obtain word representations for Russian language and APD is finally applied [23, 27]. Reference [113] claims that the approach is useful to overstep the lack of lexicographic supervision for low-resource languages and that most concept definitions in English also hold in other languages, such as Russian. However, this claim is not completely satisfied, since some words can drastically change their meaning across languages. For example, the Russian word "пионер" (i.e., pioneer, scout) is strongly connected to the Communist ideology in the Soviet Period, but it is not in the English language.

Average pairwise distance between sense prototypes (APDP). This measure is an extension of APD, and it considers all the pairs of sense prototypes $c_{1,i}$ and $c_{2,i}$ instead of all the original embeddings in Φ_1 and Φ_2 [66]. The *average pairwise distance between sense prototypes (APDP)* is defined as:

$$APD(\Psi_1, \Psi_2) = \frac{1}{|\Psi_2||\Psi_2|} \cdot \sum_{c_{1,k} \in \Psi_1,\ c_{2,k} \in \Psi_2} d(c_{1,k}, c_{2,k}). \tag{16}$$

Wassertein distance (WD). This measure models the change assessment as an *optimal transport problem,* and it is exploited as an alternative to cluster alignment when aggregation by clustering is performed separately over the embeddings Φ_1 and Φ_2 [100]. WD quantifies the effort of re-configuring the cluster distribution of p_1 into p_2, namely, minimizing the cost of moving one unit of mass (i.e., a sense prototype) from Ψ_1 to Ψ_2. The *Wassertein distance* (WD) is defined as:

$$WD(p_1, p_2) = \min_{\gamma} \sum_{i}^{k_1} \sum_{j}^{k_2} CD(c_{1,i}, c_{2,j})\ \gamma_{c_{1,i} \to c_{2,j}} \tag{17}$$

$$\text{such that:} \quad \gamma_{c_{1,i} \to c_{2,j}} \geq 0$$

$$\sum_{i} \gamma_{c_{1,i} \to c_{2,j}} = p_1$$

$$\sum_{j} \gamma_{c_{1,i} \to c_{2,j}} = p_2,$$

where all $\gamma_{c_{1,i} \to c_{2,j}}$ represents the (unknown) effort required to reconfigure the mass distribution p_1 into p_2; k_1 and k_2 are the number of clusters obtained by clustering Φ_1 and Φ_2, respectively; CD is the cosine distance computed over the sense prototypes $c_{1,i} \in \Psi_1$ and $c_{2,j} \in \Psi_2$ [17].

4.3 Ensemble-based Approaches

In this section, we review the approaches that rely on an *ensemble mechanism*, namely, the combination of two or more assessment functions to determine the semantic change score. Ensembling can mean that more than one form- and/or sense-based measure is adopted in a given approach. Ensembling can also mean that a disciplined use of both static and large LMs is used. A final semantic change score is then returned by the whole ensemble process.

According to Table 5, we note that all the ensemble approaches are time-oblivious with the exception of References [110] and [117]. We also note that unsupervised learning modalities are

Table 5. Summary View of Ensemble Approaches

Ref.	Time awareness	Learning modality	Language model	Training language	Type of training	Layer	Layer aggregation	Clustering algorithm	Change function	Corpus language
Pömsl and Lyapin 2020	time-aware	unsupervised	BERT-base, mBERT-base	monolingual, multilingual	fine-tuned	last	–	–	APD	English, German, Latin, Swedish
Teodorescu et al. 2022	time-oblivious	unsupervised	XLM-large	multilingual	trained	last four	sum	–	APD	Spanish
Martinc et al. 2020c	time-oblivious	unsupervised	BERT-base, mBERT-base	monolingual, multilingual	domain-adaptation	last four	sum	AP	CD, JSD	English, German, Latin, Swedish
Wang et al. 2020	time-oblivious	unsupervised	mBERT-base	multilingual	pre-trained	last	–	GMMs, K-means	APD, HD, JSD	Italian
Giulianelli et al. 2022	time-oblivious	unsupervised	XLM-R-base	multilingual	domain-adaptation	all	average	–	APD, PRT	English, German, Italian, Latin, Norwegian, Russian, Swedish
Ryzhova et al. 2021	time-oblivious	unsupervised	ELMo, RuBERT	monolingual, multilingual	pre-trained trained	–	–	–	APD	Russian
Kutuzov et al. 2022b	time-oblivious	unsupervised	BERT-base, ELMo	monolingual, multilingual	domain adaptation	last	–	–	APD, PRT	English, German, Latin, Swedish
Rachinskiy and Arefyev 2021	time-oblivious	supervised	XLM-R-base	multilingual	fine-tuned, pre-trained	–	–	–	APD	Russian
Rosin and Radinsky 2022	time-aware	unsupervised	BERT-base	monolingual	fine-tuned	–	–	–	CD	English, Latin, German

Missing information is denoted with a dash.

adopted with the exception of Reference [113]. As a further remark, most of the ensemble solutions exploit LLMs trained over different languages.

Some ensemble approaches combine form-based and sense-based measures to improve the quality of results. On the one hand, form-based measures are exploited to better capture the dominant sense of the target word w. On the other hand, sense-based measures are exploited to represent all the meanings of w, including the minor ones. The combination of CD (see form-based approaches in Section 4.1) and JSD (see sense-based approaches in Section 4.2) is proposed in Reference [93]. As a further ensemble experiment, the results of combining APD, HD, and JSD are discussed in Reference [138]. The APD measure is also considered in Reference [113], where multiple change scores are calculated by using different distance metrics (e.g., Manatthan distance, CD, Euclidean distance), and these scores are exploited to train a regression model as an ensemble.

Ensemble approaches based on two form-based measures are also proposed. For instance, in Reference [46], the final semantic change s is obtained by averaging APD and PRT scores. This is motivated by experimental results where sometimes APD outperforms PRT, while some other times PRT outperforms APD [73].

Some other ensemble approaches are based on the idea to combine static and contextualized embeddings. The intuition is that static embeddings can capture the dominant sense of the target word w better than form-based, contextualized embeddings. In References [110, 134], the semantic change s is assessed by leveraging both static and contextualized embeddings. In particular, s is determined by the linear combination of the scores obtained by two approaches: (i) the APD measure over contextualized embeddings (see form-based approaches in Section 4.1); (ii) the CD measure over static embeddings aligned according to the approach described in Reference [53]. Similarly, in Reference [93], instead of directly using the APD measure, JSD is exploited over clusters of contextualized embeddings (see sense-based approaches in Section 4.2). As a further difference, the scores obtained by static and contextualized approaches are combined by multiplication. The intuition is that, since the score distributions of the two approaches are unknown, multiplication prevents an approach from contributing more than the other one in the final score.

Approaches can also be combined with grammatical profiles under the intuition that grammatical changes are slow and gradual, while lexical contexts can change very quickly [46, 77]. Grammatical profile vectors gp_1 and gp_2 are associated with the times t_1 and t_2, respectively, to represent morphological and syntactical features of the considered language in the time period. In Reference [122], the contextualized embeddings of the word w occurrences are combined with the grammatical vectors. A linear regression model with regularization is trained by using as features the cosine similarities over Φ_1 and Φ_2 and over the grammatical vectors gp_1 and gp_2.

As a further ensemble approach, the combination of different time-aware techniques such as temporal attention and time masking was tested by Reference [117] to better incorporate time into word embeddings.

4.4 Discussion

According to Sections 4.1–4.3, we note that form-based approaches are more popular than sense-based ones. Most papers are characterized by time-oblivious approaches, and only a few time-aware approaches have recently appeared (e.g., Reference [117]). All approaches leverage unsupervised learning modalities with few exceptions (e.g., Reference [58]). We argue that the motivation is due to the recent introduction of a reference evaluation framework for semantic change assessment proposed at SemEval-2020 Shared Task 1, where participants were asked to adopt an unsupervised configuration [125].

All papers are featured by contextualized word embeddings extracted from BERT-like models. Regardless of their version (i.e., tiny, small, base, large), BERT and XLM-R are the most frequently used LLMs, and only a few experiments rely on ELMo and RoBERTa. As a matter of fact, the size of data needed to train or fine-tune an XLM-R model is several orders of magnitude greater than BERT. Moreover, even if less frequently employed than BERT, ELMo seems to be promising for LSC and outperforms BERT, while being much faster in training and inference [73]. As a further interesting remark, the use of static *document* embeddings extracted from a Doc2Vec[84] model has been proposed to provide pseudo-contextualized *word* embeddings as an alternative to BERT [105].

Monolingual and multilingual LLMs are both popular. The BERT models are the most frequently used monolingual models. XLM-R models are generally preferred to **mBERT (multilingual BERT)** models, since the former are trained on a larger amount of data and languages, thus the intuition is that they can better encode the language usages. Multilingual models are used both in multilingual settings, where corpora of different languages are considered (e.g., Reference [91]), and monolingual settings, where just corpora of one language are given (e.g., in Reference [46]). In a monolingual setting, the use of a multilingual model is motivated by two reasons: (i) a model pre-trained on a specific language is not available (e.g., Reference [73]), (ii) multilingual models are employed to exploit their cross-lingual transferability property (e.g., Reference [113]).

Considering the type of training, most of the papers directly use pre-trained LLMs or fine-tune them for domain adaptation. Only a few papers propose to exploit a specific fine-tuning (e.g., Reference [110]) or to incrementally fine-tune a pre-trained LLM (e.g., Reference [73]). Experiments indicate that fine-tuning a pre-trained LLM for domain adaptation consistently boosts the quality of results when compared against pre-trained LLMs (e.g., Reference [112]). The impact of fine-tuning on performance is analyzed in Reference [92], where it is shown that optimal results are achieved by fine-tuning a pre-trained LLM for five epochs and that, after five epochs, performance decreases due to overfitting. However, we argue that the fine-tuning effectiveness strictly depends on the size and domain of the considered corpora. In many papers, a different number of epochs is proposed with varying results (e.g., Reference [73]).

When an LLM is used, contextualized word embeddings are typically extracted from the last one or the last four layers of the model. Experiments show that the semantic features of text are

mainly encoded in the last four encoder layers of BERT [33, 62]. In some papers, contextualized embeddings are extracted by aggregating the output of the first and the last encoded layers. In this case, the idea is to combine *surface* features (i.e., phrase-level information, [62]) encoded in the first layer with the semantic features from the last one. Only in Reference [81] is the standalone use of lower layers of BERT proposed. Middle layers of BERT are usually excluded, since they mainly encode syntactic features [62]. When contextualized embeddings are extracted from more than one layer, they are generally aggregated by average or sum (e.g., Reference [105]). As an alternative, the use of concatenation is proposed in Reference [64].

As a further note, when an LLM is used, some words may be split into word pieces by a subword-based tokenization algorithm [129, 140]. In this case, word piece representations are generally synthesized into a single word representation $e_{j,k}$ through averaging (e.g., Reference [91]) or concatenating (e.g., Reference [93]). As alternative to avoid such problem, the pre-trained vocabulary associated with the LLM can be extended by adding some words of interest. Then, a fine-tuning step is performed to learn the weights associated with the added words (e.g., Reference [116]).

Clustering operations are typically exploited in sense-based approaches to perform Word Sense Induction [1, 4, 83, 90]. The only form-based approach that relies on clustering is presented in Reference [12] (see Section 4.1 for details). The clustering algorithms most frequently employed are K-means and **Affinity Propagation (AP)**. Further considered clustering algorithms are **Gaussian Mixture Models (GMMs)** (e.g., Reference [118]), **agglomerative clustering (AGG)** (e.g., Reference [7]), DBSCAN (e.g., Reference [65]), HDBSCAN (e.g., Reference [118]), **Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH)** (e.g., Reference [118]), **A Posteriori affinity Propagation (APP)** (e.g., Reference [105]), and **Incremental Affinity Propagation based on Nearest neighbor Assignment (IAPNA)** (e.g., Reference [105]). Since K-means, GMMs, and AGG require to define the number of clusters in advance, the use of a silhouette score is generally employed to determine the optimal number of clusters [120]. As an alternative, the AP algorithm is employed to let emerge the number of clusters without prefixing it. DBSCAN is proposed due to its capability of reducing noise by specifying (i) the minimum number of embeddings of each cluster and (ii) the maximum distance ϵ between two embeddings in a cluster. HDBSCAN is the hierarchical version of DBSCAN, and it can manage clusters of different sizes. As a difference with DBSCAN, HDBSCAN can detect noise without the ϵ parameter. APP and IAPNA are incremental extensions of AP, and their use is proposed for LSC when more than one time interval is considered. In Reference [118], different clustering algorithms are compared and the experiments show that (i) DBSCAN is very sensitive to scale, since ϵ is predefined, and (ii) BIRCH tends to find a lot of small clusters that are marginal with respect to word meanings.

Considering the change functions, a detailed presentation of possible alternatives has been provided in Sections 4.1 and 4.2. As a final remark, we note that CD and APD are frequently exploited in form-based approaches, while JSD is commonly employed in sense-based approaches.

Finally, as for the language of considered corpora, most papers consider the shared benchmark datasets taken from competitive evaluation campaigns (e.g., LSCDiscovery, [143]). Common considered languages are English, German, Latin, and Swedish that appeared in 2020 at SemEval Task 1 [125]. Russian appeared in 2021 at RuShiftEval [75, 76]. Spanish appeared in 2022 at LSCDiscovery [143]. The Italian language was introduced in 2020 at DIACRIta [11]. The approach described in Reference [91] represents a novel attempt to consider a diachronic corpus containing texts of different languages, namely, English and Slovenian.

5 Comparison of Approaches on Performances

In this section, we propose a comparison of the reviewed approaches based on their performance, considering the evaluation framework adopted in LSC tasks of shared competitions.

The framework is based on a reference benchmark that contains a diachronic textual corpus in a given language. The framework is also characterized by a test-set of target words, where each word is associated with a continuous change score (i.e., *gold score*), typically calculated based on manual annotation following the established **Word Usage Graph (WUG)** paradigm [127].[1] Different metrics are also defined in the framework to evaluate the performance of the approaches according to the kind of assessment question that the task aims to address, namely, *Grade/Binary Change*, *Sense Gain/Loss* (see Section 2).

In Table 6, we compare the reviewed approaches by considering the experiments on *Graded Change Detection* task performed and reported in the corresponding literature papers. In such a kind of task, the Spearman's correlation score is typically employed for assessing the performance of a given experiment by measuring the correlation between the predicted change scores and the gold scores.[2] The Spearman's correlation evaluates the monotonic relationship between the rank-order of the predicted scores and the gold ones. When multiple experiments are discussed in a paper, the best Spearman's correlation score obtained is reported in Table 6.

In the comparison, 12 diachronic corpora are exploited. In particular, we consider: (i) the 4 SemEval datasets [125] for English (SemEval English), German (SemEval German), Latin (SemEval Latin), and Swedish (SemEval Swedish); (ii) the English dataset proposed in Reference [51] (GEMS English); (iii) the English LiverpoolFC dataset proposed in Reference [32] (LivFC English); (iv) the COHA English dataset (COHA English); (v) the LSCDiscovery dataset [143] for Spanish (LSCD Spanish); (vi) the DURel dataset for German (DURel German) [126]; (vii) the RuShiftEval dataset for Russian (RSE Russian) [76]; and (viii) the NorDiaChange dataset for Norwegian (NOR Norwegian) [78]. In Table 6, for each corpus, we highlight when a single time interval $C_1 - C_2$ or two consecutive time intervals $C_1 - C_2$ and $C_2 - C_3$ are considered, respectively. As a further remark, we note that the RSE Russian corpus is the only case where a test-set for the time interval $C_1 - C_3$ as a whole is provided.

For the sake of readability, the performance according to the Spearman's correlation scores shown in Table 6 are labeled with the semantic change function of the considered approach and the corresponding framing with respect to form-based, sense-based, and ensemble-based categories (see Section 4).

As a general remark, we cannot find an approach outperforming all the others on all the considered corpora. This can suggest that an approach is language-dependent, namely, it works well on one language and it is not appropriate for others. By relying on the experiments presented in Reference [73], the performance of an approach is influenced by the employed assessment measure in relation with the distribution of the gold scores in the considered test-set. The experiments in Reference [73] show that when the distribution of the gold scores is skewed, namely, some words are highly changed and some others are barely changeed, the APD measure achieves better performance on Spearman's correlation than the PRT measure. On the contrary, when the distribution of the gold scores is almost uniform, namely, most of the words are similarly changed, the PRT measure achieves better performance than the APD measure.

[1] In the WUG annotation paradigm, human annotators provide semantic proximity judgments for pairs of word usages sampled from a diachronic corpus spanning two time periods. Word usages and judgments are represented as nodes and edges in a weighted, diachronic graph called *diachronic* WUG. This graph is then clustered with the correlation clustering algorithm [10], and the resulting clusters are interpreted as *word senses*. Finally, for a given word, a ground truth score of semantic change is computed by comparing the probability distributions of clusters across different time periods, e.g., a cluster with most of its usages from one time period indicates a substantial semantic change.

[2] In Reference [49], as an alternative to the Spearman's correlation score, the *Discount Cumulative Gain* is proposed. However, most papers still use Spearman's, since it is currently employed in competitive shared tasks.

Table 6. The Spearman's Correlation Score of Reviewed Approaches in Selected Experiments

Ref.	SemEval Englsh $C_1 - C_2$	SemEval German $C_1 - C_2$	SemEval Latin $C_1 - C_2$	SemEval Swedish $C_1 - C_2$	GEMS English $C_1 - C_2$	LivFC English $C_1 - C_2$	COHA English $C_1 - C_2$	LSCD Spanish $C_1 - C_2$	DURel German $C_1 - C_2$	RSE Russian $C_1 - C_2$	RSE Russian $C_2 - C_3$	RSE Russian $C_1 - C_3$	NOR Norwegian $C_1 - C_2$	NOR Norwegian $C_2 - C_3$
Teodorescu et al. 2022	–	–	–	–	–	–	–	*ensemble* APD .573	–	–	–	–	–	–
Zhou and Li 2020	*form* CD .392	*form* CD .392	*form* CD .392	*form* CD .392	–	–	–	–	–	–	–	–	–	–
Montariol et al. 2021	*sense* AP+WD .456	*sense* AP+JSD .583	*form* CD .496	*sense* K-means+WD .332	***sense*** **AP+JSD** **.510**	–	–	–	*sense* AP+JSD .712	–	–	–	–	–
Periti et al. 2022	*sense* AP+JSD .514*	–	*sense* APP+JSD .512*	–	–	–	–	–	–	–	–	–	–	–
Pömsl and Lyapin 2020	*ensemble* APD .246	*ensemble* APD .725	*ensemble* APD .463	*ensemble* APD .546	–	–	–	–	***ensemble*** **APD** **.802**	–	–	–	–	–
Rachinskiy and Arefyev 2021	–	–	–	–	–	–	–	–	–	*ensemble* APD .781	*ensemble* APD .803	*ensemble* APD .822	–	–
Rachinskiy and Arefyev 2022	–	–	–	–	–	–	–	**sense APDP .745**	–	–	–	–	–	–
Rodina et al. 2020	–	–	–	–	–	–	–	–	–	*form* PRT .557	*sense* AP+JSD .406	–	–	–
Rosin et al. 2022	*form* CD .467	–	*form* CD .512	–	–	***form*** **TD** **.620**	–	–	–	–	–	–	–	–
Rosin and Radinsky 2022	***form*** **CD** **.627**	***form*** **CD** **.763**	*form* CD .565	–	–	–	–	–	–	–	–	–	–	–
Rother et al. 2020	*sense* HDBSCAN .512	*sense* GMMs .605	*sense* GMMs .321	*sense* HDBSCAN .308	–	–	–	–	–	–	–	–	–	–
Ryzhova et al. 2021	–	–	–	–	–	–	–	–	–	*ensemble* regression .480*	*ensemble* regression .487*	*ensemble* regression .560*	–	
Kudisov and Arefyev 2022	–	–	–	–	–	–	–	*form* APD .637	–	–	–	–	–	–
Kutuzov 2020	*form* APD .605	*form* PRT .740	*form* PRT .561	***form*** **APD** **.610**	*sense* AP+JSD .456*	–	–	–	–	–	–	–	–	–
Laicher et al. 2021	*form* APD .571*	*form* CD .755*	–	*form* APD .602*	–	–	–	–	–	–	–	–	–	–
Liu et al. 2021	*form* CD .341	*form* CD .512	*form* CD .304	*form* CD .304	*form* CD .286	*form* CD .561	–	–	–	–	–	–	–	–
Martinc et al. 2020c	*ensemble* AP+JSD .361	*ensemble* AP+JSD .642	*form* CD .496	*ensemble* AP+JSD .343	–	–	–	–	–	–	–	–	–	–
Giulianelli et al. 2020	–	–	–	–	*form* APD .285*	–	–	–	–	–	–	–	–	–
Giulianelli et al. 2022	*form* APD .514	*ensemble* PRT .354	***ensemble*** **PRT** **.572**	*ensemble* APD .397	–	–	–	–	–	*ensemble* APD+PRT .376	*form* APD .480	*form* APD .457	***ensemble*** **APD+PRT** **.394**	***ensemble*** **APD** **.503**
Hu et al. 2019	–	–	–	–	–	–	***sense*** **MNS** **.428***	–	–	–	–	–	–	–
Kanjirangat et al. 2020	*sense* K-means+JSD .028*	*sense* K-means+JSD .173*	*sense* K-means+JSD .253*	*sense* K-means+CSC .321*	–	–	–	–	–	–	–	–	–	–
Karnysheva and Schwarz 2020	*sense* K-means+JSD -.155*	*sense* DBSCAN+JSD .388*	*sense* DBSCAN+JSD .177*	*sense* K-means+JSD -.062*	–	–	–	–	–	–	–	–	–	–
Kashleva et al. 2022	–	–	–	–	–	–	–	*sense* APDP .553	–	–	–	–	–	–
Keidar et al. 2022	*form* APD .489	–	–	–	–	–	–	–	–	–	–	–	–	–
Arefyev et al. 2021	–	–	–	–	–	–	–	–	–	***form*** **APD** **.825**	***form*** **APD** **.821**	***form*** **APD** **.823**	–	–
Arefyev and Zhikov 2020	*sense* AGG+CD .299	*sense* AGG+CD .094	*sense* AGG+CD −.134	*sense* AGG+CD .274	–	–	–	–	–	–	–	–	–	–
Beck 2020	*form* CD .293*	*form* CD .414*	*form* CD .343*	*form* CD .300*	–	–	–	–	–	–	–	–	–	–
Cuba Gyllensten et al. 2020	*form* CD .209*	*form* CD .656*	*form* CD .399*	*form* CD .234*	–	–	–	–	–	–	–	–	–	–
Kutuzov et al. 2022b	*form* APD .605	*form* PRT .740	*form* PRT .561	*form* APD .569	*form* APD .394	–	–	–	–	–	–	–	–	–

For each corpus, the top performance is reported in bold. Asterisks denote experiments based on a pre-trained model.

As a further remark, we note that the approaches characterized by fine-tuning achieve greater performance. This is also confirmed in the experiments of Reference [92] where fine-tuning an LLM boosts the performance when the LLM is not affected by under- or overfitting.

On average, form-based approaches outperform sense-based approaches in Graded Change Detection tasks. We argue that such a result is motivated by the structure of the test-sets, where

just one semantic change score is provided for each target word. Form-based approaches benefit from this structure, since they work on measuring the change over one general word property (i.e., the dominant sense, or the degree of polysemy). On the opposite, sense-based approaches are disadvantaged by this structure, since they work on measuring the change over multiple word meanings, and they need to produce a single, comprehensive change value that summarizes all the single-meaning changes for the comparison against the gold score. As a result, capturing some (minor) meanings can negatively affect the comprehensive change value, and to address this issue, small clusters are usually considered as possible noise and filtered out [93].

Table 6 shows that form-based approaches based on APD, CD, or PRT measures tend to obtain higher performance than sense- and ensemble-based approaches. GEMS English, COHA English, and LSCD Spanish are the only benchmarks where sense-based approaches outperform form-based ones. This can be motivated by the small number of experiments performed. Indeed, for COHA English, experiments with form-based approaches have not been tested [58], while only a few experiments and a limited number of configurations with form-based approaches have been tested on GEMS English. For LSCD Spanish, the top performance is .745, and the corresponding approach leverages the APDP measure, which is an extension of APD characterized by the use of an average-of-average operation. This result is in line with the intuition presented in Reference [105], where the use of averaging on top of clustering contributes to reduce the noise in the contextualized embeddings of the target word.

We also note that ensemble approaches are, on average, characterized by high performance. In particular, top performances are provided by ensemble approaches on SemEval Latin (.572), DURel German (.802), and NOR Norwegian (.394 and .503). Notably, the performances on SemEval Latin are obtained by combining contextualized embeddings and grammatical profiles, thereby confirming that word meanings are influenced by morphology and syntax, especially in some languages. It is also interesting to observe that the performances on DURel German are obtained through an approach combining static and contextualized word embeddings, thus highlighting that such a kind of combination can be effective. For NOR Norwegian in the time interval $C_1 - C_2$, the best approach exploits both APD and PRT; this is a further confirmation that APD and PRT are top-performing measures in semantic change detection. For the subsequent time interval $C_2 - C_3$, the best result on NOR Norwegian is obtained with a combination of APD with grammatical profiles. This is a confirmation of the intuition presented in Reference [46], which suggests that ensembling grammatical profiles with contextualized embeddings can enhance performance by incorporating morphological and syntactic features not fully captured by LLMs.

For SemEval English, SemEval German, the top performances are .627, .763, respectively, and they are obtained by the time-aware approach proposed in Reference [117]. Also, for LivFC English (.620), the top performance is obtained by leveraging a time-aware approach [116]. We argue that extra-linguistic information (e.g., time information) can have a positive impact on performance. The injection of extra-linguistic information can contribute to increase the performance also when small-size LLMs are employed, since they are less affected by noise than larger models. As a confirmation, in contrast to the widespread belief that the larger the models, the higher the performance, the best result for SemEval English is obtained by exploiting contextualized embeddings generated from a BERT-tiny model [117, 136]. This is also true for SemEval Swedish (.610), where the top performance is obtained by calculating the APD measure over contextualized embeddings extracted from an ELMo model [72], which is far smaller than LLMs.

Finally, we note that also the use of supervised learning modalities contributes to achieve high performance. As an example, the top performances for RSE Russian are .825 on $C_1 - C_2$, .821 on $C_2 - C_3$, and .823 on $C_1 - C_3$ and they are obtained by a form-based, supervised approach [6]. This

is also confirmed by the recent introduction of a novel LLM called XL-LEXEME [21], which has demonstrated exceptional performance across multiple benchmarks [107].

6 Scalability, Interpretability, and Robustness Issues

In this section, we analyze the LSC approaches by considering possible scalability, interpretability, and reliability issues.

6.1 Scalability Issues

In the LSC approaches, any occurrence of the target word considered for change assessment is represented by a specific embedding. As a basic implementation, all the contextualized embeddings are stored in memory for processing. The higher the number of occurrences of a target word, the higher the number of embeddings to manage. As a result, when the size of the diachronic corpus grows, possible issues arise both in terms of memory and computation time. Similar issues occur when multiple target words are considered for change assessment. In this case, a possible workaround for addressing the memory issue is to process one target word at a time. However, in this way, the memory issue *changes* to a computation time issue. For feasibility convenience, most experiments work on a small set of target words. This kind of limitation inhibits the possibility to address tasks like the detection of the most changed word in a corpus. The need to work on solutions capable of dealing with such a kind of scalability issue has recently been promoted in LSCDiscovery, where participants were asked to assess the semantic change on all the words of the dictionary [143].

Some possible solutions to the scalability issues have been proposed in the literature. For instance, approaches based on measures that enforce aggregation by averaging (e.g., CD, PRT) are time-scalable, since only the prototypes are considered for change assessment instead of the whole set of embeddings. Also, approaches based on APD or JSD measures can be adjusted to become time-scalable. In particular, the number of embeddings to store and process can be reduced by random sampling the occurrences of the target word w. This means that (i) a smaller number of similarity scores needs to be calculated with APD (e.g., Reference [122]), and (ii) JSD works on top of clustering algorithms that converge faster (e.g., Reference [115]). As an alternative to random sampling, an online *aggregation by summing* method is proposed in Reference [100], where a predefined number of contextualized embeddings n is stored in memory. An embedding e is stored when the number of embeddings in memory is less than n and e is strongly dissimilar from all the other embeddings previously stored. If e is not stored, then it is aggregated to the most similar embedding stored in memory through sum.

The dimensionality reduction of the embeddings is proposed as a further alternative to enforce scalability. For example, in Reference [118], the embedding dimensionality is reduced to 10 (from 768) by combining an autoencoder with the **UMAP (Uniform Manifold Approximation and Projection)** algorithm [94]. In Reference [67], UMAP and PCA are used to project contextualized embedding into $h \in \{2, 5, 10, 20, 50, 100\}$ dimensions. With respect to this solution, we argue that, although it can improve the memory scalability, time scalability is negatively affected, since dimensionality reduction takes time. However, in Reference [118], it is shown that the dimensionality reduction can still contribute to time scalability when the goal is to test and compare the effectiveness of different clustering algorithms and the reduced embeddings are saved and re-used. As a further option, the use of small LLMs, such as TinyBert or ELMo, is gaining more attention, since the dimension of the generated embeddings is far lower (e.g., Reference [117]).

Scalability issues can also arise when the change needs to be assessed on a corpus $C = \bigcup_i^n C_i$ defined over more than one time interval ($n > 2$). Typically, existing approaches calculate the change score s over each pair of time intervals (t_i, t_{i+1}) by iteratively re-applying the same

assessment workflow. As a difference, an incremental approach based on a clustering algorithm called ***A Posteriori affinity Propagation*** **(APP)** is proposed in References [22, 105, 106] to speed up the aggregation stage. In each time interval, clustering is incrementally executed by considering the prototypes of the previous time period (i.e., aggregation by averaging) and the incoming embeddings of the current time period.

6.2 Interpretability Issues

Interpretability issues arise when it is not possible to determine which meaning(s) have changed among all the meanings of a target word, namely, the meaning(s) that mainly caused the change score assessed by a considered approach. Definitely, form-based approaches are affected by such a kind of issue, since they model the change as the change in the dominant sense or in the degree of polysemy of a word without considering the possible multiple meanings. On the opposite, sense-based approaches aim at providing an interpretation of the word change, since they attempt to model the change by considering the multiple word senses. However, interpretability issues can arise also when sense-based approaches are employed due to three main motivations.

Word meaning representation. Sense-based approaches mostly rely on clustering techniques to represent word meanings. The K-means and the AP clustering algorithms are usually employed to this end. K-means requires that the number of target clusters is predefined, and this can be inappropriate to effectively represent the meanings of a target word that are not known beforehand. AP lets the number of target clusters emerge, but experimental results show that the association of a cluster with a word meaning can be imprecise. We argue that this can be due to the distributional nature of LLMs that tends to capture changes in contextual variance (i.e., word usages) rather than changes in lexicographic senses (i.e., word meanings) [79]. As an example, sometimes AP produces more than 100 clusters, which is rather unrealistic if we assume that a cluster represents a word meaning [105]. As a matter of fact, a word may completely change its context without changing its meaning [92].

Word meaning description. Each cluster obtained during the aggregation stage of a sense-based approach needs to be associated with a description that denotes the corresponding word meaning. This can be done by human experts on the basis of the cluster contents. However, this is time-consuming, given that a cluster can consist of several hundreds/thousands of elements. As an alternative, clustering analysis techniques have been proposed to label clusters by summarizing their contents. As a possible option, a cluster description can be extracted from the content by considering the top featuring keywords based on lexical occurrences (e.g., TF-IDF) [68, 100] or substitutes [20]. In Reference [45], the sense-prototype of a cluster is proposed as a cluster exemplar and the corpus sentences that are closest to the prototype are adopted as cluster/meaning description. However, when a cluster contains outliers, these sentences could not provide an effective description. More recently, the use of Causal LLMs has been proposed to generate descriptive cluster interpretations [22] or word usage definitions [47].

Word meaning evolution. When a corpus $C = \bigcup_i^n$ defined over more than one time interval is considered, the clusters defined at a timestep t_i need to be linked to the clusters of the previous timestep t_{i-1} to trace the evolution of the corresponding meaning over time (i.e., cluster/meaning history). Since the clustering executions at each timestep are independent, the capability of recognizing corresponding clusters/meanings at different timesteps can be challenging. As a possible solution, alignment techniques can be employed to link similar word meanings in different, consecutive time periods [64, 100]. As a further option, evolutionary clustering algorithms can be exploited without requiring any alignment mechanism across time periods [22, 105, 106].

6.3 Robustness Issues

Robustness issues arise when the assessment score is not reliable due to data imbalance, model stability, and model bias.

Data imbalance. The diachronic corpus C must equally reflect the presence of the target word w in both the timesteps t_1 and t_2. This means that the frequency of w must not strongly change in the considered time period. However, in common scenarios, more documents are available for the most recent timestep t_2 and *"it may not be possible to achieve balance in the sense expected from a modern corpus"* [131]. As a consequence, the frequency of w can be strongly higher in t_2 than in t_1, and the embeddings Φ_j can produce a distorted representation of the target word when the LLM is trained/fine-tuned (e.g., References [139, 146]). As a further remark, data imbalance issues can occur when some word meanings are more frequent than others. For instance, the dominant sense is usually more represented than other senses in the corpus C. As a result, when a sense-based approach is adopted, the embedding distributions p_1, p_2 can be skewed, meaning that a larger number of embeddings is associated with the dominant sense rather than with the other minor senses. In sense-based approaches, the word meanings are represented by clusters, and *the number of clusters consistently reflects word frequency* [72]. When a meaning is associated with a few embeddings/clusters, its contribution to the overall assessment score is marginally leading to an inflated or underestimated assessment score. In this respect, a qualitative analysis of "potentially erroneous" outputs of reviewed approaches is presented in Reference [79]. Some examples of potentially erroneous assessment scores occur when (i) a *word with strongly context-dependent meanings* is considered whose embeddings are mutually different; (ii) a *word is frequently used in a very specific context* in only one timestep t_1 or t_2; (iii) a *word is affected by a syntactic change*, not a semantic one. In Reference [86], a solution is proposed to reduce the false discovery rate and to improve the precision of the change assessment by leveraging permutation-based statistical test and term-frequency thresholding.

Model stability. Pre-trained LLMs are usually trained on modern text sources. For example, the original English BERT model is pre-trained on Wikipedia and BooksCorpus [147]. As a result, pre-trained LLMs are prone to represent words from a modern perspective, and thus they tend to ignore the temporal information of a considered corpus. This way, when historical corpora are considered, the possible obsolete word usages cannot be properly represented. This problem has been investigated in the literature by comparing the performance of pre-trained against fine-tuned LLMs [73, 112]. In line with the considerations of Section 4.4, the results show that fine-tuning the LLM on the whole diachronic corpus improves the quality of word representations for historical texts. Since fine-tuning the LLM can be expensive in terms of time and computational resources, a measure for estimating the model effectiveness for historical sources is presented in Reference [61]. In particular, this measure is used to decide whether a model should be re-trained or fine-tuned.

Model bias. Contextualized embeddings can possibly be affected by biases on the encoded information. For instance, a possible bias can arise from orthographic information, such as the word form and the position of a word in a sentence, since they influence the output of the top BERT layers [81]. Text pre-processing techniques are proposed as a solution to reduce the influence of orthography in the embeddings, thus increasing the robustness of encoded semantic information. To this end, lower-casing the corpus text is a commonly employed solution. However, *the lower-casing of words often conflates parts of speech*, thus another possible bias can arise. For example, the proper noun `Apple` and the common noun `apple` become identical after lower-casing [54]. The possible bias introduced by Named Entities and proper nouns is investigated in References [81, 93]. In Reference [112], text normalization techniques are proposed based on the removal of accent

markers. In some languages, such a kind of normalization can introduce a bias, since different words can be conflated. For example, papà (e.g., the Italian word for dad) and papa (e.g., the Italian word for pope) cannot be distinguished after the accent removal. Further text pre-processing techniques can be employed to reduce the possible bias due to orthographic information. In Reference [125], lemmatization and punctuation removal are proposed. Experimental results on lemmatization for reducing the model bias on BERT embeddings are presented in Reference [81]. Further experiments show that lemmatizing the target word alone is more beneficial than lemmatizing the whole corpus [81]. Filtering out content-light words, such as stop words and low-frequency words, can also be beneficial [145]. As an alternative solution to reduce word-form biases, the embedding of a word occurrence can be computed by averaging its original embedding and the embeddings of its nearest words in the input sentence [145].

When aggregation by clustering is enforced, the possible word-form biases can affect the clustering result [81]. As a solution, clustering refinement techniques have been proposed. As an option, the removal of the clusters containing only one or two instances is adopted, since they are not considered significant [93]. As a further option, in Reference [92], clusters with less than two members are considered as weak clusters, and they are merged with the closest strong cluster, i.e., cluster with more than two members. In Reference [105], clusters containing less than 5% of the whole set of embeddings are assumed to be poorly informative and are thus dropped. However, we argue that the use of clustering refinement techniques must be carefully considered, since also small clusters can be important when the corpus is unbalanced in the number of meanings of a word.

7 Challenges and Concluding Remarks

In this article, we analyzed the LSC task by providing a formal definition of the problem and a reference classification framework based on meaning representation, time awareness, and learning modality dimensions. The literature approaches are surveyed according to the given framework by considering the assessment function, the employed LLM, the achieved performance, and the possible scalability/interpretability/robustness issues.

While we provide a solid framework for classification LSC approaches, we acknowledge that the NLP research on semantic change is rapidly evolving with new papers continually emerging. For example, various models such as LLaMA [103], GPT [107], and ChatGPT [104] are being considered for LSC. Approaches based on lexical substitutes are gaining popularity to analyze both the modern and the historical bias of LLMs [30]. Further, supervised [133] and unsupervised [2] approaches, along with different change functions [3], are appearing. Additionally, new benchmarks for a larger gamma of languages are becoming available, including Chinese [24, 25], Japanese [85], and Slovenian [111].

In References [54, 74], an overview of open challenges for LSC is presented. In the following, we extend such an overview by focusing on those challenges that are specific to the existing approaches in relation with the issues discussed in Section 6.

Scalability. The trend in LSC is to adopt increasingly larger models with the idea that they better represent language features. As a consequence, scalability issues arise, and they are being addressed as discussed in Section 6.1. However, contrary to this trend, we argue that the use of small-size models, such as those introduced in References [116, 117], needs to be further explored, since they are competitive in terms of performance.

Word meaning representation. In Section 5, we show that form-based approaches outperform sense-based approaches in the Graded Change Detection assessment. However, we argue that sense-based approaches are promising, since they focus on encoding word senses, and they can enrich the mere degree of semantic change of a word w with the information about the specific meaning of w that changed. In this direction, LSC should be considered as a temporal/diachronic

extension of other problems such as Word Sense Induction [5], Word Meaning Disambiguation [48], and Word-in-Context [88].

So far, word senses have been represented through aggregation by clustering under the idea that each cluster represents a specific word meaning. However, according to the interpretability issues of Section 6, clustering techniques are often affected by noise, and they are typically capable of representing word usages rather than word meanings. Thus, further investigations are required to represent lexicographic meanings in a more faithful way.

Word meaning description. According to Section 6, current solutions to meaning description are focused on determining a representative label taken from the cluster contents (e.g., TF-IDF, sentence(s) featuring the sense-prototype). Such solutions are mostly oriented to highlight the lexical features of the cluster/meaning without considering any element that reflects the cluster's semantics. As a consequence, open challenges are based on the need of comprehensive description techniques capable of capturing both lexical and semantic aspects, such as position in text, semantics, or co-occurrences across different documents. In a very recent work, Reference [47] proposes interpreting the meaning of word usages by generating sense definitions through novel generative models. A main drawback is that different definitions can be generated for usages related to the same meaning. Nonetheless, we strongly suggest a change towards the latter solution, given that the new generative models have demonstrated extraordinary capabilities.

Word meaning evolution. In shared competitions, the reference evaluation framework for LSC is based on one/two time periods that are considered for LSC. The extension of the evaluation framework to consider more time periods is an open challenge. In particular, methods and practices of LSC approaches need to be tested/extended for detecting both short- and long-term semantic changes as well as for promoting the design of incremental techniques able to handle dynamic corpora (i.e., corpora that become progressively available).

In this context, a further challenge is about the capability to trace the change of a meaning over multiple timesteps (i.e., meaning evolution). As mentioned in Section 2, alignment techniques can be used to link similar word meanings in different, consecutive time periods. However, such a solution is not completely satisfactory due to possible limitations (e.g., scalability, robustness of alignment), and further research work is needed to better track the meaning evolution over time (e.g., Reference [105]).

Model stability. Most of the approaches surveyed in this article are time-oblivious and face the problem of model stability through fine-tuning. Since this practice can be expensive in terms of time and resources, we argue that further research on the development of time-aware approaches is needed, in that they do not suffer the model stability problem.

Model bias. The solutions to model bias issues presented in Section 6 are language-dependent, and they are mainly exploited in approaches based on monolingual models. Further research work is needed to test the effectiveness of existing solutions also in approaches based on multilingual LLMs. In addition, we argue that future work should concern the application of denoising and debiasing techniques to both monolingual and multilingual LLMs (e.g., Reference [63]) with the aim to improve LSC performance by reducing orthographic biases regardless of the language(s) on which the models were trained.

Further challenges not strictly related to the issues of Section 6 are the following:

Semantic Change Interpretation. Most of the literature papers do not investigate the nature of the detected change, meaning that they do not classify the semantic change according to the existing linguistics theory (e.g., amelioration, pejoration, broadening, narrowing, metaphorization, metonymization, and metonymy) [19, 55]. Further studies on the causes and types of semantic changes are needed [31]. These studies could be crucial to detect "laws" of semantic change that describe the condition under which the meanings of words are prone to change. For example, some

laws are hypothesized in References [35, 53, 142], but later the validity of some of them has been questioned [36]. Contextualized embeddings could contribute to test the validity of current laws and to propose new ones. To the best of our knowledge, some steps in this direction are only moved in Reference [58] for modeling the word change from an ecological viewpoint.

Computational models of meaning change. Almost all experiments on LSC are based on BERT embeddings. Although there are open questions about how to maximize the effectiveness of BERT embeddings in different language setups, the effectiveness of BERT for LSC has been extensively investigated. We believe that LSC should be extended by considering a wider range of models. Some work explored the effectiveness of ELMo [73, 115]. However, the performance of ELMo in different contexts and setups should be analyzed in more detail. Furthermore, it might be worth investigating smaller versions of BERT, such as ALBERT [82] and DistilBERT [123]. Further models can also be considered such as seq2seq and generative models, which recently showed interesting results in the field of temporal Word-in-Context problem [89].

Multilingual models. In past shared competitions on LSC, monolingual models have generally been preferred to multilingual ones. We believe that a systematic comparison of monolingual vs. multilingual models is required to determine scenarios and conditions where the former type of models provides better performance than the latter type or vice versa. Multilingual embeddings can also contribute to LSC, since they could enable a language-independent semantic change assessment, meaning that the gold-scores of different languages can be exploited as a whole for the evaluation of a given approach.

Cross-language change detection. As introduced in Reference [91], further investigations are required to address the problem of cross-language change detection. We argue that solutions to such a kind of problem can also be useful for LSC, since they can detect semantic change of *cognates* and *borrowings* (e.g., Reference [41]), as well as *contact-induced* semantic changes (e.g., Reference [97]).[3]

Use cases. So far, LSC through contextualized embeddings is still a theoretical problem not yet integrated in real application scenarios such as historical information retrieval, lexicography, linguistic research, or social-analysis. Among the existing use cases, semantic change has been examined by Reference [16] to investigate sudden events that radically alter public opinion on a topic, and by References [95, 102] to explore shifts in olfactory perception and changes in the descriptions of smells over time. We expect that further use cases and experiences will be developed and shared in the next future.

Context change over different domains. The attention gained by diachronic semantic change detection through the use of word embeddings paved the way for modeling other linguistics issues such as the identification of diatopic lexical variation [128], the detection of semantic changes of grammatical constructions [40], or the comparison of how speakers who disagree on a subject use the same words [43]. The reviewed approaches can be tested and possibly extended to cope with such a kind of linguistics issue.

Acknowledgments

We would like to thank Nina Tahmasebi for her valuable comments and constructive feedback on the manuscript. We would also like to thank the anonymous reviewers for their helpful comments, as well as research colleagues who reached out to provide feedback and suggestions on our arXiv pre-print [99].

[3]In linguistics, cognates are sets of words in different languages that have been inherited in direct descent from an etymological ancestor in a common parent language. Borrowings (or loanwords) are words adopted by the speakers of one language from a different language. Contact-induced semantic changes are diachronic changes within a recipient language that are traceable to languages other than the direct ancestor of the recipient language and that have spread and are conventionalized within a community speaking the recipient language.

References

[1] Eneko Agirre and Aitor Soroa. 2007. SemEval-2007 Task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval'07)*, Eneko Agirre, Lluís Màrquez, and Richard Wicentowski (Eds.). Association for Computational Linguistics, 7–12. Retrieved from https://aclanthology.org/S07-1002

[2] Taichi Aida and Danushka Bollegala. 2023. Swap and predict—Predicting the semantic changes in words across corpora by context swapping. In *Findings of the Association for Computational Linguistics (EMNLP'23)*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 7753–7772. DOI: https://doi.org/10.18653/v1/2023.findings-emnlp.520

[3] Taichi Aida and Danushka Bollegala. 2024. A Semantic Distance Metric Learning approach for Lexical Semantic Change Detection. DOI: https://doi.org/10.48550/arXiv.2403.00226

[4] Anna Aksenova, Ekaterina Gavrishina, Elisei Rykov, and Andrey Kutuzov. 2022. RuDSI: Graph-based word sense induction dataset for Russian. In *Proceedings of the Conference on Graph-based Methods for Natural Language Processing (TextGraphs'22)*, Dmitry Ustalov, Yanjun Gao, Alexander Panchenko, Marco Valentino, Mokanarangan Thayaparan, Thien Huu Nguyen, Gerald Penn, Arti Ramesh, and Abhik Jana (Eds.). Association for Computational Linguistics, 77–88. Retrieved from https://aclanthology.org/2022.textgraphs-1.9

[5] 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'20)*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 3171–3180. DOI: https://doi.org/10.18653/v1/2020.findings-emnlp.284

[6] Nikolay Arefyev, Maksim Fedoseev, Vitaly Protastov, Daniil Homiskiy, Adis Davletov, and Alexander Panchenko. 2021. DeepMistake: Which senses are hard to distinguish for a word-in-context model. In *Proceedings of the Conference on Computational Linguistics and Intellectual Technologies (Dialogue'21)*. RSUH. DOI: https://doi.org/10.28995/2075-7182-2021-20-16-30

[7] 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 14th Workshop on Semantic Evaluation*, Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova (Eds.). International Committee for Computational Linguistics, 171–179. DOI: https://doi.org/10.18653/v1/2020.semeval-1.20

[8] Anthony Baez and Horacio Saggion. 2023. LSLlama: Fine-tuned LLaMA for lexical simplification. In *Proceedings of the 2nd Workshop on Text Simplification, Accessibility and Readability*, Sanja Štajner, Horacio Saggio, Matthew Shardlow, and Fernando Alva-Manchego (Eds.). INCOMA Ltd., Shoumen, Bulgaria, 102–108. Retrieved from https://aclanthology.org/2023.tsar-1.10

[9] David Bamman and Patrick J. Burns. 2020. Latin BERT: A Contextual Language Model for Classical Philology. DOI: https://doi.org/10.48550/arXiv.2009.10053

[10] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. 2004. Correlation clustering. *Mach. Learn.* 56 (2004), 89–113. DOI: https://doi.org/10.1023/B:MACH.0000033116.57574.95

[11] 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'20)*. CEUR-WS. Retrieved from https://ceur-ws.org/Vol-2765/paper158.pdf

[12] Christin Beck. 2020. DiaSense at SemEval-2020 Task 1: Modeling sense change via pre-trained BERT embeddings. In *Proceedings of the 14th Workshop on Semantic Evaluation*, Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova (Eds.). International Committee for Computational Linguistics, 50–58. DOI: https://doi.org/10.18653/v1/2020.semeval-1.4

[13] Andreas Blank. 1997. *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen*. Max Niemeyer Verlag, Berlin, Boston. DOI: https://doi.org/10.1515/9783110931600

[14] Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, 1006–1017. DOI: https://doi.org/10.18653/v1/2020.acl-main.95

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

[16] Brian Bonafilia, Bastiaan Bruinsma, Denitsa Saynova, and Moa Johansson. 2023. Sudden semantic shifts in Swedish NATO discourse. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, Vishakh Padmakumar, Gisela Vallejo, and Yao Fu (Eds.). Association for Computational Linguistics, 184–193. DOI: https://doi.org/10.18653/v1/2023.acl-srw.28

[17] Nicolas Bonneel, Michiel van de Panne, Sylvain Paris, and Wolfgang Heidrich. 2011. Displacement interpolation using Lagrangian mass transport. *ACM Trans. Graph.* 30, 6 (Dec. 2011), 1–12. DOI: https://doi.org/10.1145/2070781.2024192

[18] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20)*. Curran Associates Inc., Red Hook, NY, Article 159, 25 pages. DOI: https://doi.org/doi/abs/10.5555/3495724.3495883

[19] Lyle Campbell. 2020. *Historical Linguistics*. Edinburgh University Press, Edinburgh. DOI: https://doi.org/9781474463133

[20] 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)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 590–602. DOI: https://doi.org/10.18653/v1/2023.acl-short.52

[21] 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)*. Association for Computational Linguistics, 1577–1585. DOI: https://doi.org/10.18653/v1/2023.acl-short.135

[22] Silvana Castano, Alfio Ferrara, Stefano Montanelli, and Francesco Periti. 2024. Incremental Affinity Propagation based on Cluster Consolidation and Stratification. DOI: https://doi.org/10.48550/arXiv.2401.14439

[23] Ming-Wei Chang, Lev Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2 (AAAI'08)*. AAAI Press, 830–835.

[24] Jing Chen, Emmanuele Chersoni, and Chu-ren Huang. 2022. Lexicon of changes: Towards the evaluation of diachronic semantic shift in Chinese. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, Simon Hengchen, Haim Dubossarsky, and Lars Borin (Eds.). Association for Computational Linguistics, 113–118. DOI: https://doi.org/10.18653/v1/2022.lchange-1.11

[25] 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*, Nina Tahmasebi, Syrielle Montariol, Haim Dubossarsky, Andrey Kutuzov, Simon Hengchen, David Alfter, Francesco Periti, and Pierluigi Cassotti (Eds.). Association for Computational Linguistics, 93–99. Retrieved from https://aclanthology.org/2023.lchange-1.10

[26] Ting-Rui Chiang and Dani Yogatama. 2023. The distributional hypothesis does not fully explain the benefits of masked language model pretraining. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 10305–10321. DOI: https://doi.org/10.18653/v1/2023.emnlp-main.637

[27] Hyunjin Choi, Judong Kim, Seongho Joe, Seungjai Min, and Youngjune Gwon. 2021. Analyzing zero-shot cross-lingual transfer in supervised NLP tasks. In *Proceedings of the 25th International Conference on Pattern Recognition (ICPR'21)*. 9608–9613. DOI: https://doi.org/10.1109/ICPR48806.2021.9412570

[28] Paul Cook, Jey Han Lau, Diana McCarthy, and Timothy Baldwin. 2014. Novel word-sense identification. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers (COLING '14)*, Junichi Tsujii and Jan Hajic (Eds.). Dublin City University and Association for Computational Linguistics, 1624–1635. Retrieved from https://aclanthology.org/C14-1154

[29] Amaru Cuba Gyllensten, Evangelia Gogoulou, Ariel Ekgren, and Magnus Sahlgren. 2020. SenseCluster at SemEval-2020 Task 1: Unsupervised lexical semantic change detection. In *Proceedings of the 14th Workshop on Semantic Evaluation*, Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova (Eds.). International Committee for Computational Linguistics, 112–118. DOI: https://doi.org/10.18653/v1/2020.semeval-1.12

[30] Miriam Cuscito, Alfio Ferrara, and Martin Ruskov. 2024. How BERT Speaks Shakespearean English? Evaluating Historical Bias in Contextual Language Models. DOI: https://doi.org/10.48550/arXiv.2402.05034

[31] Jader Martins Camboim de Sá, Marcos Da Silveira, and Cédric Pruski. 2024. Survey in Characterization of Semantic Change. DOI: https://doi.org/10.48550/arXiv.2402.19088

[32] Marco Del Tredici, Raquel Fernández, and Gemma Boleda. 2019. Short-term meaning shift: A distributional exploration. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 2069–2075. DOI: https://doi.org/10.18653/v1/N19-1210

[33] 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 Conference of the North American Chapter*

of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. DOI: https://doi.org/10.18653/v1/N19-1423

[34] Haim Dubossarsky, 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*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 457–470. DOI: https://doi.org/10.18653/v1/P19-1044

[35] Haim Dubossarsky, Yulia Tsvetkov, Chris Dyer, and Eitan Grossman. 2015. A bottom up approach to category mapping and meaning change. In *Proceedings of the NetWordS Final Conference*. CEUR-WS, 66–70. Retrieved from https://ceur-ws.org/Vol-1347/paper14.pdf

[36] Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, 1136–1145. DOI: https://doi.org/10.18653/v1/D17-1118

[37] Alessio Ferrari, Beatrice Donati, and Stefania Gnesi. 2017. Detecting domain-specific ambiguities: An NLP approach based on Wikipedia crawling and word embeddings. In *Proceedings of the IEEE 25th International Requirements Engineering Conference Workshops (REW'17)*. 393–399. DOI: https://doi.org/10.1109/REW.2017.20

[38] Tony Finch. 2009. Incremental calculation of weighted mean and variance. *Univ. Cambridge* 4, 11-5 (2009), 41–42. Retrieved from https://fanf2.user.srcf.net/hermes/doc/antiforgery/stats.pdf

[39] John Rupert Firth. 1957. A synopsis of linguistic theory. *Stud. Ling. Anal.* (1957).

[40] Lauren Fonteyn, F. Karsdorp, B. McGillivray, A. Nerghens, and M. Wevers. 2020. What about grammar? Using BERT embeddings to explore functional-semantic shifts of semi-lexical and grammatical constructions. In *Proceedings of the Workshop on Computational Humanities Research (CHR'20)*. CEUR-WS, 257–268. Retrieved from https://ceur-ws.org/Vol-2723/short15.pdf

[41] Clémentine Fourrier and Syrielle Montariol. 2022. Caveats of measuring semantic change of cognates and borrowings using multilingual word embeddings. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, Simon Hengchen, Haim Dubossarsky, and Lars Borin (Eds.). Association for Computational Linguistics, 97–112. DOI: https://doi.org/10.18653/v1/2022.lchange-1.10

[42] Aina Garí Soler and Marianna Apidianaki. 2021. Let's play mono-poly: BERT can reveal words' polysemy level and partitionability into senses. *Trans. Assoc. Computat. Ling.* 9 (2021), 825–844. DOI: https://doi.org/10.1162/tacl_a_00400

[43] Aina Garí Soler, Matthieu Labeau, and Chloé Clavel. 2022. One word, two sides: Traces of stance in contextualized word representations. In *Proceedings of the 29th International Conference on Computational Linguistics*, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, 3950–3959. Retrieved from https://aclanthology.org/2022.coling-1.347

[44] Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang. 2023. *Lexical Variation and Change: A Distributional Semantic Approach*. Oxford University Press, United Kingdom. DOI: https://doi.org/10.1093/oso/9780198890676.001.0001

[45] 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*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, 3960–3973. DOI: https://doi.org/10.18653/v1/2020.acl-main.365

[46] Mario Giulianelli, 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*, Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, Simon Hengchen, Haim Dubossarsky, and Lars Borin (Eds.). Association for Computational Linguistics, 54–67. DOI: https://doi.org/10.18653/v1/2022.lchange-1.6

[47] 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)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 3130–3148. DOI: https://doi.org/10.18653/v1/2023.acl-long.176

[48] Mihir Godbole, Parth Dandavate, and Aditya Kane. 2022. Temporal word meaning disambiguation using TimeLMs. In *Proceedings of the the 1st Workshop on Ever Evolving NLP (EvoNLP'22)*, Francesco Barbieri, Jose Camacho-Collados, Bhuwan Dhingra, Luis Espinosa-Anke, Elena Gribovskaya, Angeliki Lazaridou, Daniel Loureiro, and Leonardo Neves (Eds.). Association for Computational Linguistics, 55–60. DOI: https://doi.org/10.18653/v1/2022.evonlp-1.8

[49] 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*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, 538–555. DOI: https://doi.org/10.18653/v1/2020.acl-main.51

[50] Roksana Goworek and Haim Dubossarsky. 2024. Toward sentiment aware semantic change analysis. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, Neele Falk, Sara Papi, and Mike Zhang (Eds.). Association for Computational Linguistics, 350–357. Retrieved from https://aclanthology.org/2024.eacl-srw.28

[51] Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the Workshop on GEometrical Models of Natural Language Semantics (GEMS'11)*, Sebastian Pado and Yves Peirsman (Eds.). Association for Computational Linguistics, 67–71. Retrieved from https://aclanthology.org/W11-2508

[52] Helena Halmari. 2011. Political correctness, euphemism, and language change: The case of "People First." *J. Pragmat.* 43, 3 (2011), 828–840. DOI: https://doi.org/10.1016/j.pragma.2010.09.016

[53] 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)*, Katrin Erk and Noah A. Smith (Eds.). Association for Computational Linguistics, 1489–1501. DOI: https://doi.org/10.18653/v1/P16-1141

[54] Simon Hengchen, Nina Tahmasebi, Dominik Schlechtweg, and Haim Dubossarsky. 2021. *Challenges for Computational Lexical Semantic Change*. Language Science Press, Berlin, 341–372. DOI: https://doi.org/10.5281/zenodo.5040322

[55] Hans Henrich Hock and Brian D. Joseph. 2019. *Language History, Language Change, and Language Relationship*. De Gruyter Mouton, Berlin, Boston. DOI: https://doi.org/10.1515/9783110613285

[56] 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)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 6970–6984. DOI: https://doi.org/10.18653/v1/2021.acl-long.542

[57] Franziska Horn. 2021. Exploring word usage change with continuously evolving embeddings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, Heng Ji, Jong C. Park, and Rui Xia (Eds.). Association for Computational Linguistics, 290–297. DOI: https://doi.org/10.18653/v1/2021.acl-demo.35

[58] Renfen Hu, 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*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 3899–3908. DOI: https://doi.org/10.18653/v1/P19-1379

[59] Han Huang, Tomoyuki Kajiwara, and Yuki Arase. 2021. Definition modelling for appropriate specificity. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 2499–2509. DOI: https://doi.org/10.18653/v1/2021.emnlp-main.194

[60] Xiaolei Huang and Michael J. Paul. 2019. Neural temporality adaptation for document classification: Diachronic word embeddings and domain adaptation models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 4113–4123. DOI: https://doi.org/10.18653/v1/P19-1403

[61] Shotaro Ishihara, Hiromu Takahashi, and Hono Shirai. 2022. Semantic shift stability: Efficient way to detect performance degradation of word embeddings and pre-trained language models. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (Eds.). Association for Computational Linguistics, 205–216. Retrieved from https://aclanthology.org/2022.aacl-main.17

[62] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 3651–3657. DOI: https://doi.org/10.18653/v1/P19-1356

[63] Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, 1256–1266. DOI: https://doi.org/10.18653/v1/2021.eacl-main.107

[64] 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 14th Workshop on Semantic Evaluation*, Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova (Eds.). International Committee for Computational Linguistics, 214–221. DOI: https://doi.org/10.18653/v1/2020.semeval-1.26

[65] Anna Karnysheva and Pia Schwarz. 2020. TUE at SemEval-2020 Task 1: Detecting semantic change by clustering contextual word embeddings. In *Proceedings of the 14th Workshop on Semantic Evaluation*, Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova (Eds.). International Committee for Computational Linguistics, 232–238. DOI: https://doi.org/10.18653/v1/2020.semeval-1.28

[66] Kseniia Kashleva, Alexander Shein, Elizaveta Tukhtina, and Svetlana Vydrina. 2022. HSE at LSCDiscovery in Spanish: Clustering and profiling for lexical semantic change discovery. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, Simon Hengchen, Haim Dubossarsky, and Lars Borin (Eds.). Association for Computational Linguistics, 193–197. DOI: https://doi.org/10.18653/v1/2022.lchange-1.21

[67] 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)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 1422–1442. DOI: https://doi.org/10.18653/v1/2022.acl-long.101

[68] 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*, Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, Simon Hengchen, Haim Dubossarsky, and Lars Borin (Eds.). Association for Computational Linguistics, 131–139. DOI: https://doi.org/10.18653/v1/2022.lchange-1.14

[69] Cunliang Kong, Yun Chen, Hengyuan Zhang, Liner Yang, and Erhong Yang. 2022. Multitasking framework for unsupervised simple definition generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 5934–5943. DOI: https://doi.org/10.18653/v1/2022.acl-long.409

[70] 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*, Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, Simon Hengchen, Haim Dubossarsky, and Lars Borin (Eds.). Association for Computational Linguistics, 165–172. DOI: https://doi.org/10.18653/v1/2022.lchange-1.17

[71] Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for Russian language. In *Proceedings of the Conference on Computational Linguistics and Intellectual Technologies (Dialogue'19)*. RSUH. DOI: https://doi.org/10.48550/arXiv.1905.07213

[72] Andrey Kutuzov. 2020. *Distributional Word Embeddings in Modeling Diachronic Semantic Change*. Ph. D. Dissertation. University of Oslo. Retrieved from https://www.duo.uio.no/handle/10852/81045

[73] Andrey Kutuzov and Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 Task 1: Contextualised embeddings for lexical semantic change detection. In *Proceedings of the 14th Workshop on Semantic Evaluation*, Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova (Eds.). International Committee for Computational Linguistics, 126–134. DOI: https://doi.org/10.18653/v1/2020.semeval-1.14

[74] 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*, Emily M. Bender, Leon Derczynski, and Pierre Isabelle (Eds.). Association for Computational Linguistics, 1384–1397. Retrieved from https://aclanthology.org/C18-1117

[75] Andrey Kutuzov and Lidia Pivovarova. 2021. RuShiftEval: A shared task on semantic shift detection for Russian. In *Proceedings of the Conference on Computational Linguistics and Intellectual Technologies (Dialogue'21)*. RSUH. Retrieved from https://www.dialog-21.ru/media/5536/pivovarovalpluskutuzova151.pdf

[76] 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*, Nina Tahmasebi, Adam Jatowt, Yang Xu, Simon Hengchen, Syrielle Montariol, and Haim Dubossarsky (Eds.). Association for Computational Linguistics, 7–13. DOI: https://doi.org/10.18653/v1/2021.lchange-1.2

[77] Andrey Kutuzov, Lidia Pivovarova, and Mario Giulianelli. 2021. Grammatical profiling for semantic change detection. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, Arianna Bisazza and Omri Abend (Eds.). Association for Computational Linguistics, 423–434. DOI: https://doi.org/10.18653/v1/2021.conll-1.33

[78] Andrey Kutuzov, Samia Touileb, Petter Mæhlum, Tita Enstad, and Alexandra Wittemann. 2022. NorDiaChange: Diachronic semantic change dataset for Norwegian. In *Proceedings of the 13th Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck,

Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, 2563–2572. Retrieved from https://aclanthology.org/2022.lrec-1.274

[79] Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2022. Contextualized embeddings for semantic change detection: Lessons learned. *North. Eur. J. Lang. Technol.* 8 (2022), Leon Derczynski (Ed.). Northern European Association of Language Technology. DOI: https://doi.org/10.3384/nejlt.2000-1533.2022.3478

[80] Severin Laicher, Gioia Baldissin, Enrique Castañeda, Dominik Schlechtweg, and Sabine Schulte. 2020. CL-IMS @ DIACR-Ita: Volente o Nolente: BERT does not Outperform SGNS on semantic change detection. In *Proceedings of the Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA'20)*. CEUR-WS, 438–443. Retrieved from https://ceur-ws.org/Vol-2765/paper132.pdf

[81] Severin Laicher, Sinan Kurtyigit, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Explaining and improving BERT performance on lexical semantic change detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, Ionut-Teodor Sorodoc, Madhumita Sushil, Ece Takmaz, and Eneko Agirre (Eds.). Association for Computational Linguistics, 192–202. DOI: https://doi.org/10.18653/v1/2021.eacl-srw.25

[82] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. DOI: https://doi.org/10.48550/arXiv.1909.11942

[83] Jey Han Lau, 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*, Walter Daelemans (Ed.). Association for Computational Linguistics, 591–601. Retrieved from https://aclanthology.org/E12-1060

[84] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (Proceedings of Machine Learning Research,* Vol. 32*)*, Eric P. Xing and Tony Jebara (Eds.). PMLR, 1188–1196. Retrieved from https://proceedings.mlr.press/v32/le14.html

[85] 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*, Chu-Ren Huang, Yasunari Harada, Jong-Bok Kim, Si Chen, Yu-Yin Hsu, Emmanuele Chersoni, Pranav A., Winnie Huiheng Zeng, Bo Peng, Yuxi Li, and Junlin Li (Eds.). Association for Computational Linguistics, 125–136. Retrieved from https://aclanthology.org/2023.paclic-1.13

[86] 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*, Yang Gao, Steffen Eger, Wei Zhao, Piyawat Lertvittayakumjorn, and Marina Fomicheva (Eds.). Association for Computational Linguistics, 104–113. DOI: https://doi.org/10.18653/v1/2021.eval4nlp-1.11

[87] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. DOI: https://doi.org/10.48550/arXiv.1907.11692

[88] Daniel Loureiro, Aminette D'Souza, Areej Nasser Muhajab, Isabella A. White, Gabriel Wong, Luis Espinosa-Anke, Leonardo Neves, Francesco Barbieri, and Jose Camacho-Collados. 2022. TempoWiC: An evaluation benchmark for detecting meaning shift in social media. In *Proceedings of the 29th International Conference on Computational Linguistics*, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, 3353–3359. Retrieved from https://aclanthology.org/2022.coling-1.296

[89] Chenyang Lyu, Yongxin Zhou, and Tianbo Ji. 2022. MLLabs-LIG at TempoWiC 2022: A generative approach for examining temporal meaning shift. In *Proceedings of the the 1st Workshop on Ever Evolving NLP (EvoNLP'22)*, Francesco Barbieri, Jose Camacho-Collados, Bhuwan Dhingra, Luis Espinosa-Anke, Elena Gribovskaya, Angeliki Lazaridou, Daniel Loureiro, and Leonardo Neves (Eds.). Association for Computational Linguistics, 1–6. DOI: https://doi.org/10.18653/v1/2022.evonlp-1.1

[90] Suresh Manandhar, Ioannis Klapaftis, Dmitriy Dligach, and Sameer Pradhan. 2010. SemEval-2010 Task 14: Word sense induction & disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, Katrin Erk and Carlo Strapparava (Eds.). Association for Computational Linguistics, 63–68. Retrieved from https://aclanthology.org/S10-1011

[91] Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2020. Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of the 12th Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, 4811–4819. Retrieved from https://aclanthology.org/2020.lrec-1.592

[92] Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020. Capturing evolution in word usage: Just add more clusters? In *Proceedings of the Web Conference (WWW'20)*. Association for Computing Machinery, 343–349. DOI: https://doi.org/10.1145/3366424.3382186

[93] Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020. Discovery team at SemEval-2020 Task 1: Context-sensitive embeddings not always better than static for semantic change detection. In *Proceedings of the 14th Workshop on Semantic Evaluation*, Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova (Eds.). International Committee for Computational Linguistics, 67–73. DOI: https://doi.org/10.18653/v1/2020.semeval-1.6

[94] Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. DOI: https://doi.org/10.48550/arXiv.1802.03426

[95] Stefano Menini, Teresa Paccosi, Sara Tonelli, Marieke Van Erp, Inger Leemans, Pasquale Lisena, Raphael Troncy, William Tullett, Ali Hürriyetoğlu, Ger Dijkstra, Femke Gordijn, Elias Jürgens, Josephine Koopman, Aron Ouwerkerk, Sanne Steen, Inna Novalija, Janez Brank, Dunja Mladenic, and Anja Zidar. 2022. A multilingual benchmark to capture olfactory situations over time. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, Simon Hengchen, Haim Dubossarsky, and Lars Borin (Eds.). Association for Computational Linguistics, 1–10. DOI: https://doi.org/10.18653/v1/2022.lchange-1.1

[96] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. DOI: https://doi.org/10.48550/arXiv.1301.3781

[97] Filip Miletic, Anne Przewozny-Desriaux, and Ludovic Tanguy. 2021. Detecting contact-induced semantic shifts: What can embedding-based methods do in practice? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 10852–10865. DOI: https://doi.org/10.18653/v1/2021.emnlp-main.847

[98] George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey*. Retrieved from https://aclanthology.org/H94-1111

[99] Stefano Montanelli and Francesco Periti. 2023. A Survey on Contextualised Semantic Shift Detection. DOI: https://doi.org/10.48550/arXiv.2304.01666

[100] Syrielle Montariol, Matej Martinc, and Lidia Pivovarova. 2021. Scalable and interpretable semantic change detection. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, 4642–4652. DOI: https://doi.org/10.18653/v1/2021.naacl-main.369

[101] Ben O'Neill. 2011. A critique of politically correct language. *Indep. Rev.* 16, 2 (2011), 279–291. Retrieved from http://www.jstor.org/stable/24563157

[102] Teresa Paccosi, Stefano Menini, Elisa Leonardelli, Ilaria Barzon, and Sara Tonelli. 2023. Scent and sensibility: Perception shifts in the olfactory domain. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, Nina Tahmasebi, Syrielle Montariol, Haim Dubossarsky, Andrey Kutuzov, Simon Hengchen, David Alfter, Francesco Periti, and Pierluigi Cassotti (Eds.). Association for Computational Linguistics, 143–152. DOI: https://doi.org/10.18653/v1/2023.lchange-1.15

[103] Francesco Periti, 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*. Association for Computational Linguistics.

[104] Francesco Periti, Haim Dubossarsky, and Nina Tahmasebi. 2024. (Chat)GPT v BERT dawn of justice for semantic change detection. In *Findings of the Association for Computational Linguistics (EACL'24)*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, 420–436. Retrieved from https://aclanthology.org/2024.findings-eacl.29

[105] 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*, Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, Simon Hengchen, Haim Dubossarsky, and Lars Borin (Eds.). Association for Computational Linguistics, 33–43. DOI: https://doi.org/10.18653/v1/2022.lchange-1.4

[106] 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. DOI: https://doi.org/10.36227/techrxiv.24210915.v1

[107] Francesco Periti and Nina Tahmasebi. 2024. A systematic comparison of contextualized word embeddings for lexical semantic change. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics. Retrieved from https://doi.org/10.48550/arXiv.2402.12011

[108] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, 2227–2237. DOI: https://doi.org/10.18653/v1/N18-1202

[109] Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: The word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 1267–1273. DOI: https://doi.org/10.18653/v1/N19-1128

[110] Martin Pömsl and Roman Lyapin. 2020. CIRCE at SemEval-2020 Task 1: Ensembling context-free and context-dependent word representations. In *Proceedings of the 14th Workshop on Semantic Evaluation*, Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova (Eds.). International Committee for Computational Linguistics, 180–186. DOI: https://doi.org/10.18653/v1/2020.semeval-1.21

[111] Marko Pranjić, Kaja Dobrovoljc, Senja Pollak, and Matej Martinc. 2024. Semantic Change Detection for Slovene Language: A Novel Dataset and an Approach Based on Optimal Transport. DOI: https://doi.org/10.48550/arXiv.2402.16596

[112] Wenjun Qiu and Xu Yang. 2022. HistBERT: A Pre-trained Language Model for Diachronic Lexical Semantic Analysis. DOI: https://doi.org/10.48550/arXiv.2202.03612

[113] Maxim Rachinskiy and Nikolay Arefyev. 2021. Zeroshot crosslingual transfer of a gloss language model for semantic change detection. In *Proceedings of the Conference on Computational Linguistics and Intellectual Technologies (Dialogue'21)*. RSUH. DOI: https://doi.org/dx.doi.org/10.28995/2075-7182-2021-20-578-586

[114] Maxim Rachinskiy and Nikolay Arefyev. 2022. GlossReader at LSCDiscovery: Train to select a proper gloss in English—Discover lexical semantic change in Spanish. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, Nina Tahmasebi, Andrey Kutuzov, Simon Hengchen, Haim Dubossarsky, and Lars Borin (Eds.). Association for Computational Linguistics, 198–203. DOI: https://doi.org/10.18653/v1/2022.lchange-1.22

[115] Julia Rodina, Yuliya Trofimova, Andrey Kutuzov, and Ekaterina Artemova. 2020. ELMo and BERT in Semantic Change Detection for Russian. arXiv:2010.03481 [cs.CL]

[116] Guy D. Rosin, Ido Guy, and Kira Radinsky. 2022. Time masking for temporal language models. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining (WSDM'22)*. Association for Computing Machinery, 833–841. DOI: https://doi.org/10.1145/3488560.3498529

[117] Guy D. Rosin and Kira Radinsky. 2022. Temporal attention for language models. In *Findings of the Association for Computational Linguistics (NAACL'22)*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, 1498–1508. DOI: https://doi.org/10.18653/v1/2022.findings-naacl.112

[118] David Rother, Thomas Haider, and Steffen Eger. 2020. CMCE at SemEval-2020 Task 1: Clustering on manifolds of contextualized embeddings to detect historical meaning shifts. In *Proceedings of the 14th Workshop on Semantic Evaluation*, Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova (Eds.). International Committee for Computational Linguistics, 187–193. DOI: https://doi.org/10.18653/v1/2020.semeval-1.22

[119] Paul Röttger and Janet Pierrehumbert. 2021. Temporal adaptation of BERT and performance on downstream document classification: Insights from social media. In *Findings of the Association for Computational Linguistics (EMNLP'21)*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 2400–2412. DOI: https://doi.org/10.18653/v1/2021.findings-emnlp.206

[120] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53–65. DOI: https://doi.org/10.1016/0377-0427(87)90125-7

[121] Maja Rudolph and David Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of the World Wide Web Conference (WWW'18)*. International World Wide Web Conferences Steering Committee, 1003–1011. DOI: https://doi.org/10.1145/3178876.3185999

[122] Anastasiia Ryzhova, Daria Ryzhova, and Ilya Sochenkov. 2021. Detection of semantic changes in Russian nouns with distributional models and grammatical features. In *Proceedings of the Conference on Computational Linguistics and Intellectual Technologies (Dialogue'21)*. RSUH. DOI: https://doi.org/dx.doi.org/10.28995/2075-7182-2021-20-597-606

[123] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. DOI: https://doi.org/10.48550/arXiv.1910.01108

[124] Yo Sato and Kevin Heffernan. 2020. Homonym normalisation by word sense clustering: A case in Japanese. In *Proceedings of the 28th International Conference on Computational Linguistics*, Donia Scott, Nuria Bel, and Chengqing

Zong (Eds.). International Committee on Computational Linguistics, 3324–3332. DOI: https://doi.org/10.18653/v1/2020.coling-main.295

[125] 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 14th Workshop on Semantic Evaluation*, Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova (Eds.). International Committee for Computational Linguistics, 1–23. DOI: https://doi.org/10.18653/v1/2020.semeval-1.1

[126] Dominik Schlechtweg, 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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, 169–174. DOI: https://doi.org/10.18653/v1/N18-2027

[127] Dominik Schlechtweg, 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 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 7079–7091. DOI: https://doi.org/10.18653/v1/2021.emnlp-main.567

[128] Frank Seifart. 2019. *Contact-induced Change*. De Gruyter Mouton, Berlin, Boston. 13–23. DOI: https://doi.org/10.1515/9783110435351-002

[129] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Katrin Erk and Noah A. Smith (Eds.). Association for Computational Linguistics, 1715–1725. DOI: https://doi.org/10.18653/v1/P16-1162

[130] 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 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 66–76. DOI: https://doi.org/10.18653/v1/D19-1007

[131] Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. *Survey of Computational Approaches to Lexical Semantic Change Detection*. Language Science Press, Berlin, 1–91. DOI: https://doi.org/10.5281/zenodo.5040302

[132] Xuri Tang. 2018. A state-of-the-art of semantic change computation. *Nat. Lang. Eng.* 24, 5 (2018), 649–676. DOI: https://doi.org/10.1017/S1351324918000220

[133] Xiaohang Tang, Yi Zhou, Taichi Aida, Procheta Sen, and Danushka Bollegala. 2023. Can word sense distribution detect semantic changes of words? In *Findings of the Association for Computational Linguistics (EMNLP'23)*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 3575–3590. DOI: https://doi.org/10.18653/v1/2023.findings-emnlp.231

[134] Daniela Teodorescu, Spencer von der Ohe, and Grzegorz Kondrak. 2022. UAlberta at LSCDiscovery: Lexical semantic change detection via word sense disambiguation. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, Simon Hengchen, Haim Dubossarsky, and Lars Borin (Eds.). Association for Computational Linguistics, 180–186. DOI: https://doi.org/10.18653/v1/2022.lchange-1.19

[135] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-tuned Chat Models. DOI: https://doi.org/10.48550/arXiv.2307.09288

[136] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-Read Students Learn Better: On the Importance of Pre-training Compact Models. DOI: https://doi.org/10.48550/arXiv.1908.08962

[137] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, 6000–6010. DOI: https://doi.org/doi/10.5555/3295222.3295349

[138] Benyou Wang, Emanuele Di Buccio, and Massimo Melucci. 2020. University of Padova @ DIACR-Ita. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA'20)*. CEUR-WS. Retrieved from https://ceur-ws.org/Vol-2765/paper91.pdf

[139] Laura Wendlandt, Jonathan K. Kummerfeld, and Rada Mihalcea. 2018. Factors influencing the surprising instability of word embeddings. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, 2092–2102. DOI: https://doi.org/10.18653/v1/N18-1190

[140] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. DOI: https://doi.org/10.48550/arXiv.1609.08144

[141] Katherine Wysocki and Joseph R. Jenkins. 1987. Deriving word meanings through morphological generalization. *Read. Res. Quart.* 22, 1 (1987), 66–81. DOI: https://doi.org/10.2307/747721

[142] Yang Xu and Charles Kemp. 2015. A computational evaluation of two laws of semantic change. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. Cognitive Science Society. Retrieved from https://www.cs.toronto.edu/~yangxu/xu_kemp_2015_parallelchange.pdf

[143] 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*, Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, Simon Hengchen, Haim Dubossarsky, and Lars Borin (Eds.). Association for Computational Linguistics, 149–164. DOI: https://doi.org/10.18653/v1/2022.lchange-1.16

[144] Ziqian Zeng, Xin Liu, and Yangqiu Song. 2018. Biased random walk based social regularization for word embeddings. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI '18)*. International Joint Conferences on Artificial Intelligence Organization, 4560–4566. DOI: https://doi.org/10.24963/ijcai.2018/634

[145] Jinan Zhou and Jiaxin Li. 2020. TemporalTeller at SemEval-2020 Task 1: Unsupervised lexical semantic change detection with temporal referencing. In *Proceedings of the 14th Workshop on Semantic Evaluation*, Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova (Eds.). International Committee for Computational Linguistics, 222–231. DOI: https://doi.org/10.18653/v1/2020.semeval-1.27

[146] Kaitlyn Zhou, Kawin Ethayarajh, and Dan Jurafsky. 2021. Frequency-based Distortions in Contextualized Word Embeddings. DOI: https://doi.org/10.48550/arXiv.2104.08465

[147] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'15)*. 19–27. DOI: https://doi.org/10.1109/ICCV.2015.11

Received 9 January 2023; revised 23 May 2024; accepted 6 June 2024