Language Resources and Evaluation
https://doi.org/10.1007/s10579-024-09769-1

ORIGINAL PAPER

Studying word meaning evolution through incremental semantic shift detection

Francesco Periti[1] · **Sergio Picascia**[1] · **Stefano Montanelli**[1] · **Alfio Ferrara**[1] · **Nina Tahmasebi**[2]

Accepted: 5 August 2024

Abstract

The study of *semantic shift*, that is, of how words change meaning as a consequence of social practices, events and political circumstances, is relevant in Natural Language Processing, Linguistics, and Social Sciences. The increasing availability of large diachronic corpora and advance in computational semantics have accelerated the development of computational approaches to detecting such shift. In this paper, we introduce a novel approach to tracing the evolution of word meaning over time. Our analysis focuses on gradual changes in word semantics and relies on an incremental approach to semantic shift detection (SSD) called *What is Done is Done* (WiDiD). WiDiD leverages scalable and evolutionary clustering of contextualised word embeddings to detect semantic shift and capture temporal *transactions* in word meanings. Existing approaches to SSD: (a) significantly simplify the semantic shift problem to cover change between two (or a few) time points, and (b) consider the existing corpora as static. We instead treat SSD as an organic process in which word meanings evolve across tens or even hundreds of time periods as the corpus is progressively made available. This results in an extremely demanding task that entails a multitude of intricate decisions. We demonstrate the applicability of this incremental approach on a diachronic corpus of Italian parliamentary speeches spanning eighteen distinct time periods. We also evaluate its performance on seven popular labelled benchmarks for SSD across multiple languages. Empirical results show that our results are comparable to state-of-the-art approaches, while outperforming the state-of-the-art for certain languages.

Keywords Lexical semantic change · Semantic shift detection · Contextualized word embeddings · Evolutionary clustering

Extended author information available on the last page of the article

Published online: 09 September 2024

1 Introduction

Words are malleable and their meaning(s) continuously evolve, influenced by social practices, events, and political circumstances (Azarbonyad et al., 2017). An example of this phenomenon is the word |strain|, which has recently exhibited a *semantic shift* towards the "virus strain" sense due to the COVID-19 global pandemic (Montariol et al., 2021). Traditionally, linguists and other scholars in the humanities and social sciences have studied semantic shift through time-consuming manual analysis and have thus been limited in terms of the volume, genres and time that can be considered. However, the increasing availability of large diachronic corpora and advances in computational semantics have promoted the development of computational approaches to Semantic Shift Detection (SDD)[1].

A reliable computational method for capturing the change degree of a word over time and the evolution of its individual senses would be an extremely useful tool for text-based researchers like linguists, historians and lexicographers. Figure 1 shows how the word "abuse" has changed over time. This type of result can also serve as a useful NLP resource for testing large language models on their ability to correctly capture meaning in text.

In the past decade, several studies have proven that distributional word representations (i.e., word embeddings) can be effectively used to trace semantic shift (Periti & Montanelli, 2024; Tahmasebi et al., 2021; Tang, 2018; Kutuzov et al., 2018). Thus recent advances in SSD have focused on distinguishing the multiple meanings of a word by clustering its contextualised embeddings. The idea is that each cluster should denote a specific sense that can be recognised in the documents being considered.

Since SemEval-2020 (Schlechtweg et al., 2020), there is an established evaluation framework for SSD to compare the performance of various models and approaches. Due to the substantial annotation efforts required to create reliable benchmarks over multiple time periods, the research community has generally opted to create simplified benchmarks spanning over two time periods, only. Schlechtweg et al. (2020)

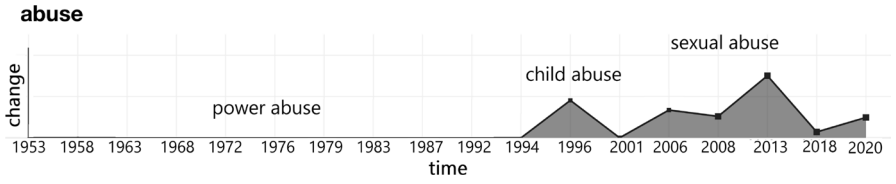

Fig. 1 Change degree of the word "abuso" (i.e., *abuse*) in a diachronic corpus of Italian parliamentary speeches and the evolution of its individual senses. Change is captured using the WiDiD approach presented in Periti et al. (2022). Before 1994, there is no change and only one sense nodule, *power abuse*; thereafter we observe changes brought about by the emergence of two more sense nodules, namely *child abuse* and *sexual abuse*

[1] Semantic shift is also often referred to as "lexical semantic change", "semantic change", as well as "sense evolution" (Bloomfield, 1933; Geeraerts, 2020).

originally provided benchmarks for English, Latin, German, and Swedish.[2] Following the success of SemEval-2020, Basile et al. (2020) introduced a benchmark for Italian,[3] Kutuzov and Pivovarova (2021) for Russian,[4] and Zamora-Reina et al. (2022) for Spanish.[5] However, thus far corpora have usually been considered in a static way, meaning that the documents are not split with respect to time period, and a single clustering activity is performed over the entire corpus. Although this generates clusters of word meanings from documents of different time periods, it does not allow us to model the full complexity of the problem. In the case of a dynamic corpus where time documents are progressively added [e.g., *posts* from social networks, (Noble et al. 2021)], capturing the evolution of multiple word meanings across tens or even hundreds of time periods represents a combinatorial explosion that vastly exceeds comparing word meanings across two time periods. To model semantic shift in a way that allows us to answer research questions posed in the humanities and social sciences, we need to model *each individual sense over all time periods*. This requires numerous comparisons, resulting in a complex and demanding task.

If the aggregation of clusters is sequentially enforced over each pair of time periods (i.e., time intervals), a set of clusters needs to be linked to the clusters of the previous time interval to trace the evolution of the corresponding meaning over time. Since the execution of clustering at each time interval is independent, alignment of corresponding meanings (i.e., clusters) at different time periods can be challenging (Kanjirangat et al., 2020; Montariol et al., 2021; Tahmasebi & Dubossarsky, 2023). To address this problem, Periti et al. (2022) recently proposed an incremental approach to SSD named *What is Done is Done* (WiDiD). This approach leverages an evolutionary and scalable clustering algorithm that facilitates direct alignment between clusters across different time periods, thereby sidestepping the need for additional cluster alignment phases. Thus far, Periti et al. (2022) have only introduced a preliminary version of the WiDiD approach. However, a thorough evaluation, application, and discussion are still warranted.

In this paper, we address these gaps by

- *extending WiDiD with cluster analysis techniques*: Originally, Periti et al. (2022) proposed WiDiD solely to quantify the degree of change experienced by a word over various time periods. In this paper, we extend the WiDiD approach by integrating cluster analysis techniques aimed at facilitating the interpretation of the detected changes. Additionally, we introduce a novel cluster visualisation method to facilitate the study of word meaning evolution over time.
- *conducting a comprehensive evaluation*: Originally, Periti et al. (2022) evaluated WiDiD against only two reference benchmarks for Latin and English on the Graded Change Detection task (Schlechtweg et al., 2020). This task consists of ranking a set of target words according to their degree of change between two time periods. In this paper, we evaluate WiDiD across multiple languages (i.e.,

[2] www.ims.uni-stuttgart.de/en/research/resources/corpora/sem-eval-ulscd/
[3] https://diacr-ita.github.io/DIACR-Ita/
[4] https://disk.yandex.ru/d/CIU9Hm0tvKPH2g
[5] https://zenodo.org/records/6433667

English, Latin, German, Swedish, Spanish, Russian, Italian) and contextualised models (i.e., BERT, mBERT, XLM-R), against seven benchmarks. Our results using WiDiD are comparable to state-of-the-art approaches, while outperforming the state-of-the-art for certain languages.

- *applying WiDiD to a real-world dataset*: To demonstrate the functionality of WiDiD, we present a case study where we apply WiDiD to a large corpus of Italian parliamentary speeches spanning eighteen different time periods (i.e., eighteen legislatures). Through this application, we engage into a detailed discussion of the implications of the WiDiD approach in supporting the study of lexical semantic change. We provide insights into both the effectiveness and limitations of WiDiD in capturing the evolution of word meanings over time and we outline future perspectives for the computational modeling of lexical semantic change, thereby contributing to the advancement of this field.

Paper structure. The remainder of the paper is organised as follows. In Sect. 2, we review the relevant literature on the use of contextualised embeddings for SSD. Our work builds on the existing WiDiD approach used for SSD (Periti et al. 2022). In particular, we extend WiDiD with novel cluster analysis to describe semantic shift and word meaning evolution. Thus, in Sect. 3, we introduce WiDiD along with the notation that will be used throughout the paper. We then present our extension in Sect. 4. A concrete application of these techniques and metrics is illustrated in Sect. 5. The results of WiDiD on the Grade Change Detection task are evaluated in Sect. 6. Finally, Sect. 7 contains our concluding remarks.

2 Related work

While approaches based on static embeddings are effective in identifying semantic shift (Tahmasebi et al., 2021; Kutuzov et al., 2018), they typically cannot differentiate the meaning(s) of a word that have remained stable from those that have changed over time. This issue has motivated recent efforts to capture word meanings using contextualised word embeddings (Periti & Montanelli, 2024). Unlike earlier approaches, approaches based on contextualised embeddings leverage a distinct word representation for each occurrence of a target word. These contextualised approaches may be either *form*-based or *sense*-based. Form-based approaches address SSD by analysing how the dominant meaning or the degree of polysemy of a word changes over time (Giulianelli et al., 2020; Martinc et al., 2020). However, like approaches based on static embeddings, they cannot differentiate the multiple meanings of a word. By contrast, sense-based approaches treat word meanings individually by enforcing clustering of contextualised embeddings (Martinc et al., 2020; Montariol et al., 2021).

Usually, all the documents for any two time periods that are being compared are available in one corpus, and a single clustering activity is performed over the entire corpus, generating clusters of word meanings from documents from the different time periods. Shift in word meaning can be detected by examining the evolution of these clusters over time. An increasing proportion of elements in a cluster indicates

that the associated word meaning is becoming more common, while a decreasing proportion suggests that the meaning is becoming obsolete. A measure of semantic shift is then employed on top of the clustering result to derive a general semantic shift assessment for a given word. For example, the cluster member distributions between two periods are often compared using the Jensen-Shannon divergence criterion (JSD) (Giulianelli et al., 2020).

Initially, Hu et al. (2019) used supervised clustering by leveraging a reference dictionary to list the possible lexicographic meanings of a word prior to analysis. However, this method relies on the availability of a digital diachronic dictionary, which is unlikely to be available for low-resource languages. Thus, a number of unsupervised clustering algorithms, like K-Means [e.g., Giulianelli et al. (2020)], HDBSCAN [e.g., Rother et al. [2020)], or Affinity Propagation [e.g., Martinc et al. (2020)] have been proposed to sidestep the need for lexicographic resources. However, unsupervised modelling of meanings without relying on external lexicographic resources tends to emphasise word usage rather than word meaning, since distributional models derive their information from the context surrounding word tokens [e.g., Kutuzov et al. (2022)]. In this case, the resulting clusters of word meanings are clusters of "**sense nodules**"—i.e., *lumps of meaning with greater stability under contextual changes* (Cruse, 2000)—rather than lexicographic meanings.

When a dynamic corpus spanning more than two time periods is considered, clusters of word meanings need to be recalculated, meaning that scalability issues arise and that the resulting clusters could change dramatically from one time period to the next. Thus, it becomes significantly more difficult to capture the possible evolutionary patterns of a word's meaning across multiple time periods. Kanjirangat et al. (2020) and Montariol et al. (2021) propose performing separate clustering activities for each time period and subsequently aligning the clustering results to recognise similar word meanings in different, consecutive time periods. However, scalability issues still arise since the clusters of word meanings need to be continuously realigned. To sidestep these issues, a promising approach called WiDiD has been proposed by Periti et al. (2022). In WiDiD, separate clustering activities are conducted using an evolutionary clustering algorithm that considers the temporal nature of the documents under consideration.

More recently, an increasing number of approaches have emerged to address SSD. Among these, *supervised* approaches and approaches based on *lexical substitutes* have gained attention. The former leverage external knowledge [e.g., dictionaries, Rachinskiy and Arefyev (2022)] or other forms of supervision [e.g., Word-in-Context datasets, Cassotti et al. (2023)] to support the shift assessment. Although they have proven to be powerful solutions against the available evaluation benchmarks, their use may not be feasible for low-resource languages or for analyzing corpora (e.g., *medical* texts from the *Middle Ages*) whose time periods and domains are not covered by the available supervision resources—a problem generally known as *temporal generalisation* (e.g., (Alkhalifa et al. 2023; Su et al. 2022)). The latter represents word senses by relying on lexical substitutes generated by a masking language model [e.g., Periti et al. (2024) and Card (2023)]. However, they typically suffer from the same limitations of current sense-based approaches, as clusters of lexical substitutes need to be aligned over time.

As SSD is typically framed in an unsupervised scenario (Schlechtweg et al., 2020), we will focus on the **WiDiD** approach, which models the time dimensions in a completely *unsupervised* way while being directly applicable to different corpora, reducing human intervention for clustering alignment. Related to WiDiD is the work by Basile et al. (2016, 2019), where an incremental approach to SSD for the Italian language is also employed. However, while they use a single vector to represent each target word in a specific time period, WiDiD relies on multiple word representations per time period.

3 WiDiD: what is done is done

WiDiD leverages an evolutionary clustering algorithm to cluster contextualised embeddings of different time periods without requiring any post hoc alignment of clusters. In WiDiD, instead of recalculating clusters at each time period, a "memory" of past word meaning clusters is maintained. In each consecutive time period, the word embeddings of that time period are compared to the already existing clusters. They either get assigned to an existing cluster or are allowed to form a new cluster, and thus the memory gets updated at each time period. As a result, the stratified layers of clusters over time allow assessment of the quantity of semantic shift as well as reconstruction of the evolution of a word's meanings.

3.1 Incremental semantic shift detection

Consider a dynamic, diachronic document corpus

$$\mathcal{C} = \bigcup_{t=0} C^t$$

where C^t denotes a set of documents added at time t. Given a target word w, our goal is to analyse how the meaning(s) of w changed along $\mathcal{C}$.

We address this problem by leveraging WiDiD. In WiDiD, documents in $\mathcal{C}$ are considered as a data stream segmented into a sequence of time periods. A four-step pipeline is repeatedly applied to the progressively added documents in $\mathcal{C}$. In Periti et al. (2022), the first three enforced steps were identified as *Document Selection* (DS), *Embedding Extraction* (EE), and *Incremental Clustering* (IC). In this paper, we extend WiDiD by enforcing an additional step of *Clustering Analysis* (CA) at the end of the pipeline (see Fig. 2).

At the first time step (i.e., $t = 0$), only the documents in C^0 are considered. As a result, only a synchronic analysis of clustering is possible, as there is no knowledge available about the meaning of w in the past. Then, for each subsequent step $t = 1...n$, the knowledge of the w meaning(s) detected in the past time periods (i.e., time periods $0...t-1$) is exploited by the IC step to cluster the documents in C^t. This diachronic analysis of clustering can provide insights into the semantic shift that has occurred.

Table 1 A reference table of notations used in the paper

Notation	Definition
w	Target word
C^t	Set of documents at time t
C^t_w	Subset of documents of C^t containing the word w
$e^t_{w,i}$	Embedding of the word w in the i-th document of C^t_w
Φ^t_w	Set of the embeddings of w in the corpus C^t_w
K^t_w	Set of clusters obtained at the t-th iteration for w
$\phi_{w,k}$	k-th cluster containing the embeddings of the word w
$\phi^t_{w,k}$	Subset of embeddings from time t in the cluster $\phi_{w,k}$
$\mu^t_{w,k}$	Prototypical representation of w for $\phi^t_{w,k}$
M^t_w	Set of prototypes $\mu^t_{w,k}$ available at time t
π^t_w	Polysemy of the word w at time t
S^t_w	Semantic shift of the word w at time t
$\rho^t_{w,k}$	Prominence of the cluster $\phi^t_{w,k}$ at time t
$\mathcal{T}^t_{w,k}$	Sense shift of the cluster $\psi_{w,k}$ at time t

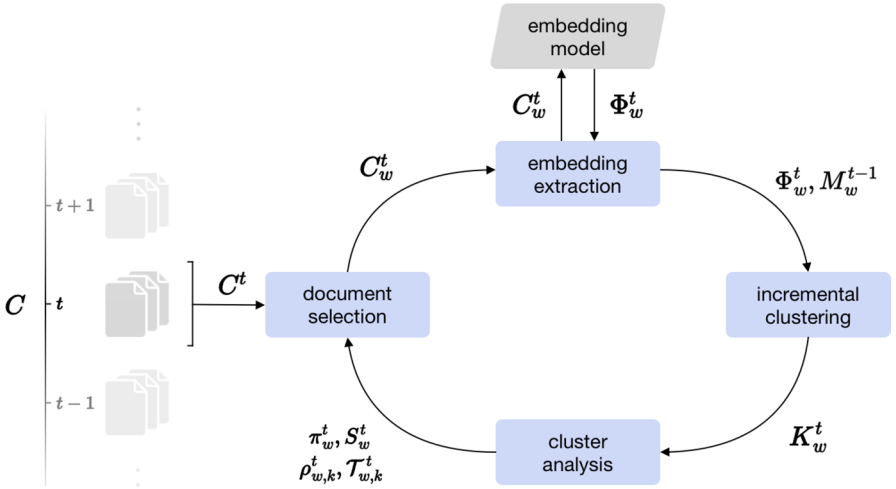

Fig. 2 WiDiD: an incremental approach to Semantic Shift Detection

The documents in C^t are processed via WiDiD as follows. For the sake of clarity, the notation used throughout this paper is summarised in Table 1.

3.1.1 Document Selection (DS)

In this step, WiDiD selects the subset of documents $C^t_w \subseteq C_t$ that contains an occurrence of the word w. Since semantic change is often accompanied by

morphosyntactic drift (Kutuzov et al., 2021), we consider any derived form of the lemma of w (e.g., plural) as an occurrence of w.

3.1.2 Embedding Extraction (EE)

In this step, WiDiD encodes each occurrence of the target word w in C_w^t with a different representation. Currently, contextualised embeddings represent the preferred tool for addressing SSD (Periti & Montanelli, 2024); thus we will use embeddings generated by standard BERT-like models (i.e., BERT, mBERT, XLM-R). However, we stress that the WiDiD approach can be employed regardless of the specific model used to represent individual word occurrences. The final output of this step is the set Φ_w^t containing all the embeddings of the word w generated for the corpus C^t. Formally,

$$\Phi_w^t = \{e_{w,1}^t, \dots, e_{w,m}^t\} ,$$

where $e_{w,j}^t$ is the contextualised embedding of w in the j-th document and m is the number of documents in C_w^t.

3.1.3 Incremental Clustering (IC)

WiDiD first ($t = 0$) uses the standard affinity propagation (AP) algorithm over Φ_w^0 (Frey & Dueck, 2007). This results in a set of clusters denoted as K_w^0.

For $t > 0$, clustering is performed using the A Posteriori affinity Propagation (APP) algorithm to cluster the embeddings Φ_w^t in groups representing different word meanings (i.e., *sense nodules*). We denote the set of resulting clusters as K_w^t.

At each time step (see Algorithm 1), APP creates an additional *sense prototype* embedding $\mu_{w,k}^{t-1}$ for each cluster $k \in K_w^{t-1}$ by averaging all its enclosed embeddings, meaning that $\mu_{w,k}^{t-1}$ is the centroid of the k-th cluster. The resulting sense prototypes constitute the "memory" of the word meanings observed so far. This memory is then exploited as the basis for subsequent word observations in the current time period. In particular, we denote as M_w^{t-1} the set of sense prototypes $\mu_{w,k}^{t-1}$ available at time $t-1$. Hence, APP consists of performing the standard AP over the set of embeddings $\Phi_w^t \cup M_w^{t-1}$. As a final step of APP, each sense prototype $\mu_{w,k}^{t-1}$ is removed, and the original embeddings compressed into $\mu_{w,k}^{t-1}$ are assigned to its corresponding cluster. This ensures that all the embeddings associated with a sense prototype at time $t-1$ are grouped together within the same cluster at the time t. This way, clusters of word meanings previously created cannot be changed (*WiDiD: What is Done is Done*), and the word meanings that are observed in the present must be stratified/integrated over the past ones. Further details are provided in Periti et al. (2022).

Notably, WiDiD represents a significantly more scalable solution than existing approaches (Montariol et al., 2021; Kanjirangat et al., 2020). Since clusters formed in previous steps are considered as unique prototypes, in each clustering step we work with a significantly smaller set of embeddings, while at the same time eliminating the need for cluster alignment techniques.

3.1.4 Clustering analysis (CA)

In this *novel* step of WiDiD, each clustering result obtained as an IC output is analysed to interpret the meaning of words from both a synchronic and diachronic perspective. This advancement of WiDiD is presented in further detail in Section 4, where we introduce a comprehensive set of metrics specifically designed to describe both a target word and its sense nodules over time.

Algorithm 1 *The APP algorithm*

```
if t == 0 then
    K_w^0 ← AP(Φ_w^0)

else
    M_w^{t-1} ← sense_prototypes(K_w^{t-1}, Φ_w^{t-1})
    K_w^t ← AP(M_w^{t-1} ∪ Φ_w^t)
    K_w^{t-1} ← sense_assignments(Φ_w^{t-1}, M_w^{t-1}, K_w^t)
end if

yield K_w^t
```

4 Cluster analysis (CA)

For each time period t, the incremental clustering (IC) results in a set of k clusters $K_w^t = \phi_{w,1}, ..., \phi_{w,k}$. In particular, we denote the set of embeddings from Φ_w^t enclosed in the k-th cluster as $\phi_{w,k}^t$. Formally, we define $\phi_{w,k}^t = \phi_{w,k} \cap \Phi_w^t$. This implies that $\phi_{w,k}^t \subset \Phi_w^t$ is the subset of embeddings extracted at time t that are members of the cluster $\phi_{w,k}$ during that specific time step.

In this paper, to be able to analyse the sequence of clustering results for a word w, we provide WiDiD with a set of metrics that characterise w both from a synchronic and diachronic perspective. Regardless of the perspective, these metrics are also conceived to inspect a particular clustering result by considering two linguistic targets:

1. *word*: when all clusters are considered overall, we analyse the target word w;
2. *sense nodules*: when a single cluster is considered, we analyse the corresponding *cluster of corpus usage* (Kutuzov et al., 2022), i.e., a sense nodule.

4.1 Synchronic perspective

From a synchronic perspective, words and sense nodules are considered within a specific time period, without taking into account their evolution in meaning. We define two metrics to describe the status of words and sense nodules, respectively.

Polysemy, denoted as π_w^t, describes the status of a word at a particular time period t. Polysemy is defined as the number of "active" sense nodules that are present at time t, i.e., sense nodules from earlier periods integrated with new elements as well as newly identified sense nodules. Intuitively, the more clusters there are, the more polysemous the word is.

$$\pi_w^t = |K_w^t| \tag{1}$$

Prominence, denoted as $\rho_{w,k}^t$, describes the status of a sense nodule at a particular time period t. Prominence is defined as the prevalence of an active sense $\phi_{w,k}^t$ at time t relative to the other active sense nodules. Intuitively, the more members in a cluster, the more prominent the sense nodule is.

$$\rho_{w,k}^t = \frac{|\phi_{w,k}^t|}{|\Phi_w^t|} \tag{2}$$

4.2 Diachronic perspective

From a diachronic perspective, words and sense nodules are considered across time periods, taking into account their evolution in meaning. The clusters at the last iteration are used in the analysis and are traced over time, thus avoiding a complex analysis of potential mergers across all time periods. We define two metrics to describe the evolution of words and sense nodules, respectively.

Semantic shift, denoted as $\mathcal{S}_w$, describes the degree of lexical semantic change of a word over two consecutive time periods. Semantic shift is defined as the degree of dissimilarity in the prominence of active sense nodules between these time periods. Intuitively, the greater the dissimilarity between time periods t and $t-1$, the higher the degree of semantic shift a word has undergone. Similar to the lexical semantic change definition in SemEval-2020 Task1 (Schlechtweg et al., 2020), $\mathcal{S}_w$ aims to capture the acquisition of a new sense nodule or the loss of an outdated sense nodule.

Following Giulianelli et al. (2020), we formally define semantic shift as the Jensen-Shannon divergence (JSD) over the prominence distributions P_w^{t-1} and P_w^t, where the $k-$th value of a distribution P_w^i is the prominence $\rho_{w,k}^i$ associated with the $k-$th sense nodule resulting from the last enforced clustering step.

$$JSD(P_w^{t-1}, P_w^t) = \frac{1}{2}\left(KL(P_w^{t-1}||M) + KL(P_w^t||M)\right),$$

where $M = (P_w^{t-1} + P_w^t)/2$, and KL represents the Kullback-Leibler divergence, as JSD is a symmetrisation of KL.

Sense shift, denoted as $\mathcal{T}_{w,k}$, describes the degree of lexical semantic change of a specific word's sense nodule over two consecutive time periods. Sense shift is defined as the degree of distance in the sense prototypes $\mu_{w,k}^t$ and $\mu_{w,k}^{t-1}$ for these time periods. Intuitively, the greater the difference between time periods t and $t-1$, the

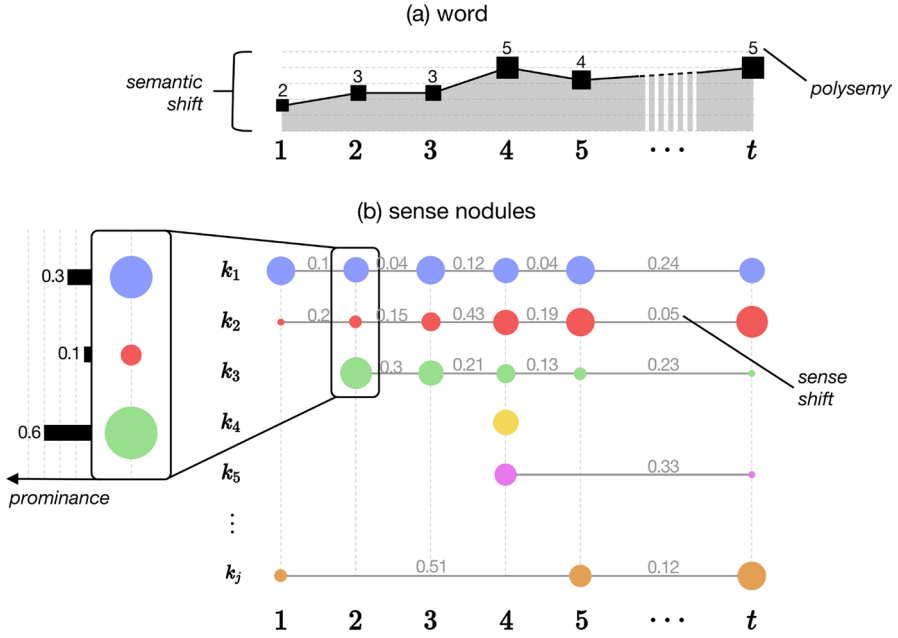

Fig. 3 Clustering visualisation: prototype visualisation of word meaning evolution. Subfigure **a** represents the polysemy and semantic shift of a word over time. Subfigure **b** represents the prominence and sense shift of the sense nodules of that word over time

greater the degree of sense shift a sense nodule undergoes. Unlike $\mathcal{S}_w$, $\mathcal{T}_{w,k}$ aims to capture lexical semantic change specific to sense nodules such as amelioration, pejoration, broadening or narrowing.

We formally define the sense shift of the $k-$th sense nodule as the cosine distance between the sense prototypes $\mu^t_{w,k}$ and $\mu^{t-1}_{w,k}$.

$$\mathcal{T}_{w,k}(\mu^t_{w,k}, \mu^{t-1}_{w,k}) = \frac{\mu^t_{w,k} \cdot \mu^{t-1}_{w,k}}{\|\mu^t_{w,k}\| \, \|\mu^{t-1}_{w,k}\|}$$

4.3 Clustering visualisation

To facilitate the analysis and interpretation of the evolution of a word's meaning, we propose a new visualisation that supports the synchronic and diachronic metrics enforced in cluster analysis. Unlike the visualisation methods for diachronic semantic shift presented in Kazi et al. (2022), this visualisation is particularly suited to a posteriori analysis of the last clustering result of WiDiD. Our visualisation provides valuable insights into the different sets of sense nodules held by a word over time, as well as clearly representing the evolution of those sense nodules.

For the sake of clarity, we describe the rationale of the visualisation by considering the prototype of an arbitrary word w illustrated in Figure 3. The figure consists

of two subfigures (a) and (b), representing the synchronic and diachronic metrics for (a) a target word and (b) its sense nodule, respectively. In both subfigures, the x -axis represents time.

In subfigure (a), each square represents a snapshot of a specific word at a particular time period t. The size of each square reflects the polysemy π_w^t of the word at time t. Semantic shift values over time are reported on the y-axis.

In subfigure (b), each circle in the figure represents a snapshot of a specific sense nodule at a particular time period t. The evolution of different sense nodules (i.e., k_1, ..., k_j) is illustrated on the y-axis using different colours. Intuitively, the presence/absence of a circle at time t indicates the active/inactive state of the related sense nodule. The size of each circle reflects the prominence ρ_w^t of the corresponding sense nodule at time t. Sense shift values over time are reported on the links connecting the snapshots of sense nodules with their respective immediately subsequent snapshots.

5 Real application of WiDiD

In this section, we report on a practical application of WiDiD involving a large corpus of Italian parliamentary speeches from 1948 to 2020. This case study is particularly relevant for detecting semantic shift as it deals with popular issues in the public and social arenas. Our main goal is to demonstrate a **practical application** of WiDiD in detecting semantic shift. Although a quantitative evaluation is not possible due to the lack of an annotated benchmark (i.e., gold scores for a set of target words), we provide a qualitative analysis of the results to assess the effectiveness of WiDiD in detecting semantic shift.

5.1 Case study dataset

Our case study dataset consists of a set of parliamentary speeches from the Italian Chamber of Deputies. It spans a period of 72 years, from the 1st legislature of the Italian Republic after the Constituent Assembly (1948) to February of the 18th Republican Legislature (2020). This dataset was created by collecting all the available plenary session transcripts at the time of downloading from the Italian Parliament website[6].

The legislatures provide a natural criterion for splitting the corpus over time, meaning that a separate sub-corpus C_i is defined for each legislature i (see 2).

[6] https://dati.camera.it/it/dati/

Table 2 Summary of the case study dataset of Italian Parliamentary speeches

	Time periods																	
Legislature	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Start date	1948	1953	1958	1963	1968	1972	1976	1979	1983	1987	1992	1994	1996	2001	2006	2008	2013	2018
End date	1953	1958	1963	1968	1972	1976	1979	1983	1987	1992	1994	1996	2001	2006	2008	2013	2018	2020
# tokens	13.0 M	13.8 M	18.3 M	18.6 M	10.1 M	8.0 M	6.0 M	11.7 M	9.6 M	11.3 M	5.2 M	4.5 M	12.8 M	12.3 M	4.3 M	12.4 M	14.3 M	5.5 M

5.2 Case study setup

To set up the case study, we first defined a set of target words whose semantic shift we would seek to detect in the Italian parliamentary corpus. Then, for each target word, we followed the WiDiD pipeline presented in Sect. 3.

Since the dataset was produced by OCR scanning, it included numerous spurious characters where words had been incorrectly recognised and introduced into the text, degrading the quality of the data. To address this issue, we performed an additional processing step to exclude speech with purely procedural content (e.g., *The MP* [SURNAME NAME] *asks to speak*) and filtered out speech associated with a high level of noise (e.g., spurious characters and other artifacts introduced during the OCR scanning process).[7] To enhance scalability in this study, as in other studies reported in the literature (Rodina et al., 2021), we reduced the number of embeddings to store and process by randomly sampling a fixed number of occurrences of each target word (i.e., 100).

We used the Transformers library by HuggingFace to extract contextual word embeddings from a pre-trained BERT model (i.e., *bert-base-multilingual-cased*[8]) without performing any fine-tuning (Wolf et al., 2020). To extract contextualised embeddings for a specific target word w, we fed the model with individual text sequences containing an occurrence of w. For each occurrence of w, we extracted a contextualised embedding from the last hidden layer of the model. Due to the byte-pair input encoding scheme employed by BERT models, some word occurrences may not correspond to words but rather to word pieces (Sennrich et al., 2016). Therefore, if a word was split into more than one sub-word, we built a single word embedding by averaging the corresponding sub-word embeddings.

Our implementation of APP was based on the original implementation released by Periti et al. (2022). The first sub-corpus (i.e., the first legislature) was considered in the initial run of AP, and then the remaining sub-corpora were added one-by-one in a specific APP iteration.

[7] Our data and code are available at https://github.com/FrancescoPeriti/WiDiD. The dataset used in our study is made available to reproduce our illustrative results. However, we decided not to release the full dataset in its current form. As discussed in the manuscript, the dataset contains a relevant number of spurious characters and OCR errors, and we are currently undertaking an extensive post-OCR cleaning process. We plan to release the dataset in the future with possible analytical insights. The cleaning process is posing considerable challenges, even with the support of advanced generative language models. While these models can help in correcting OCR errors, they tend to paraphrase or creatively reconstruct sentences Boros et al. (2024), potentially introducing artifacts that could affect the analysis of lexical semantic changes and the overall reliability of our historical, societal, and political corpus. Furthermore, we are unable to provide code for downloading data from the website (i.e., https://dati.camera.it/it/dati/) as the code is proprietary and was developed by a third-party software company under license. This restriction limits our ability to share the exact code used for data acquisition.

[8] Although we initially experimented with a monolingual pre-trained BERT model (*dbmz/bert-base-italian-uncased*), the empirical results revealed poor quality. Empirical results obtained with the multilingual model indicated a higher level of quality. We hypothesise that multilingual models can leverage their larger, cross-lingual contextualisation and pre-trained knowledge to better handle the various text quality issues present in our OCR-corrupted data.

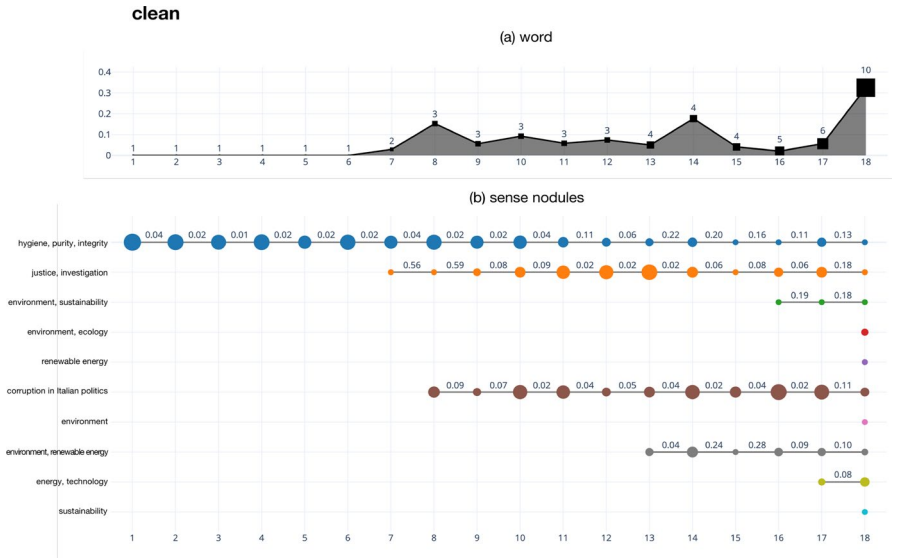

Fig. 4 Clustering visualisation: **a** *semantic shift* and *polysemy* of the Italian word "pulito" (e.g., *clean*); **b** *sense shift* and *prominence* of the sense nodules of the Italian word "pulito" (e.g., *clean*)

Manually examining sentences in a specific cluster to interpret the clusters and the semantic shift between two time periods is laborious and time-consuming. It involves a meticulous process of close-reading because multiple sentences are present within each cluster. Thus, like Montariol et al. (2021), we automatically extracted the most discriminating words for each cluster to minimise human effort. In particular, we first lemmatised each sentence within the clusters. Then, we treated each cluster as an individual document and considered all the clusters as a corpus. For each cluster, we calculated the Term Frequency-Inverse Document Frequency (TF-IDF) score of every word. To ensure the selection of the most meaningful keywords, we eliminated stopwords and excluded parts of speech other than nouns, verbs and adjectives. Thus, we obtained a ranked list of keywords for each cluster, and the top-ranked keywords were then used for cluster interpretation.

5.3 Case study results

Due to space limitations, we can provide only a few illustrative examples. However, the comprehensive list of words, including their polysemy and semantic shift as well as their sense nodules with associated prominence and sense shift, are available online for further reference .

Note that recent work has demonstrated that the geometry of BERT's embedding space exhibits anisotropy, meaning that the contextualised embeddings occupy a narrow cone within the vector space, leading to very small values of cosine

distance (Ethayarajh, 2019). Thus, for the sake of readability, we normalised the shift scores of our experiment by the maximum shift value we obtained.

As an example, Fig. 4a, b are a visual representation of the result of the cluster analysis for the Italian word |pulito| (*clean*). This word holds particular significance in the Italian context as it represents an adjective commonly associated with cleanliness. However, it gained a specific historical connotation during the early '90s owing to its association with the fight against corruption.

Figure 4a summarises Fig. 4b, providing insights into the polysemy of the word and its overall semantic shift across different time periods. The greatest semantic shifts occur in the time intervals 7–8, 13–14, and 17–18. The first time interval is associated with the acquisition of a new sense nodule (i.e., *corruption in Italian politics*). The second time interval is associated with a change in the distribution of sense nodule prominence; for example, in the 14th legislature, the sense nodule *environment, renewable energy* exhibits its maximum prominence. The third time interval is characterised by the emergence of several new sense nodules. Interestingly, the algorithm validates our expectations by capturing the emergence of new sense nodules related to the environment and renewable energy. Indeed, recent years show increasing global attention to environmental issues due to factors such as concerns about climate change.

In the discussion of Fig. 4b we adopt the ecological view of word change proposed by Hu et al. (2019). They suggest that word sense nodules can compete for dominance and cooperate for mutual benefit (i.e., remain active), similar to organisms in an ecosystem. As a complementary view of Fig. 4, Table 3 shows the proportion of documents (i.e., prominence) assigned to each sense nodule.

The cluster analysis in Fig. 4b captures examples of semantic shift of the word over time. For instance, we observe an *evergreen* sense nodule (i.e., always present across all considered time periods) associated with the label *hygiene, purity, and integrity*. This sense nodule represents the predominant meaning of the word until the 9th legislature. However, from the 10th legislature onwards, its prominence decreases due to competition with sense nodules *justice, investigation* and *corruption in Italian politics*. As with Hu et al. (2019), we find that similar senses join forces and cooperate against others while also competing internally.

On average, sense shift values are very low, indicating that sense nodules are enriched with documents that are very similar to those already existing. However, we also notice some exceptional cases with high shift scores, for example, 0.56 and 0.59 for the cluster *justice, investigation* in the time interval 7–8 and 8–9. By examining the prominence values in Table 3, we find that these cases are sometimes associated with a very small number of documents (e.g., fewer than 10 documents) rather than indicating a true sense shift, while at other times these values can be attributed to misclassification due to the quality of the considered dataset. The former observation aligns with the previous intuition by Periti et al. (2022) that computing sense prototypes of large sets of embeddings helps to reduce noise. Indeed, we observe a negative correlation between sense shift and the number of documents within a given time interval, meaning that the smaller the number of documents in a specific time interval, the more sense shift is affected by noise since the impact of outliers becomes more significant in the process of

Table 3 Prominence of the word *clean* over time

Cluster: *label*	Legislatures																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Hygiene, purity, integrity	100	72	55	70	34	60	33	58	33	36	16	10	8	12	2	4	11	2
Justice, investigation	–	–	–	–	–	–	2	1	7	17	36	44	66	18	4	11	17	1
Environment, sustainability	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	3	3	1
Environment, ecology	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	6
Renewable energy	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	3
Corruption in Italian politics	–	–	–	–	–	–	–	21	8	47	38	10	18	48	20	73	55	10
Environment	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	3
Environment, renewable energy	–	–	–	–	–	–	–	–	–	–	–	–	8	18	2	9	8	5
Energy, technology	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	6	12
Sustainability	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	3
Word frequency	100	72	55	70	34	60	35	80	48	100	90	64	100	96	28	100	100	46

Additionally, we provide the total frequency of the word over time

A dash indicates that no documents (i.e., 0) are present in that cluster at a specific time

averaging multiple embeddings (i.e. computing sense prototypes). Thus, we argue that the most significant shifts are related to medium-low sense-shift values. For example, we examined the sentences associated with cluster 0 for legislatures 11 and 12, where a sense shift of 0.11 is predicted. In the 10th legislature, the term *clean* is metaphorically used in the context of honesty, integrity, moral correctness and cleaning up criminality. The presence of comparable sentences in the 11th legislature, with a slightly different connotation emphasising the removal of corruption, old practices and dishonesty, suggests a broadening of meaning. For instance, within the 10th legislature, expressions such as "piazza pulita" (clean sweep), "mani pulite" (clean hands), "coscienza pulita" (clean conscience) are present. On the other hand, in the 11th legislature, expressions like "paese pulito" (clean country) and "ambiente pulito" (clean environment) are also present.

Further intriguing results from our analysis of various word and sense nodules are presented in Tables 4 and 5, respectively.

6 Evaluation

In this section, we evaluate the effectiveness and robustness of WiDiD by analysing its performance on various benchmarks of recent shared tasks such as SemEval-Task 1 (Schlechtweg et al., 2020), DIACRIta (Basile et al., 2020), RuShiftEval (Kutuzov & Pivovarova, 2021), and LSCDiscovery (Zamora-Reina et al., 2022). These tasks provide a rigorous evaluation framework for comparing the performance of different semantic analysis systems. The frameworks are based on a reference benchmark that contains a textual diachronic corpus in a given language. Each framework is also characterised by a test-set of target words, where each word is associated with a shift score (i.e., *gold score*) calculated on the basis of manual annotation.

To evaluate WiDiD, we rely on the Task 1 framework of SemEval-Task 1 (Schlechtweg et al., 2020), where participants are asked to solve two subtasks:

1. **Binary classification** (Subtask 1): *For a set of target words, decide which words lost or gained usage(s) between C1 and C2, and which did not*. A binary label ($l \in \{0, 1\}$) is assigned to each target word via manual annotation. Then the semantic shift word classification computed by a model is evaluated by the Accuracy over the human-annotated test data.
2. **Ranking** (Subtask 2): *Rank a set of target words according to their degree of semantic shift between C1 and C2*. A continuous score is assigned to each target word via manual annotation. Then the semantic shift word ranking computed by a model is evaluated by the Spearman's rank-order correlation over the human-annotated test data.

Originally, Periti et al. (2022) evaluated the WiDiD performance on Subtask 2 using the English and Latin corpora of SemEval. In this paper, we further evaluate WiDiD on seven different corpora. It is worth noting that the evaluation for DIACRIta was executed only on Subtask 1, since no continuous labels are provided. Conversely, the

Table 4 Example of semantic shift associated with the corresponding word, time interval, polysemy, and a short description

Word	Time-interval	Polysemy	Semantic shift	Description
Clean (*pulito*)	7–8	2–3	0.15	The term is used in the context of *corruption in Italian politics* in addition to its original associations with *hygiene, purity and integrity*.
Violence (*violenza*)	17–18	8–14	0.53	The term is used to encompass not just physical violence, sexual assault, and domestic violence, but also gender-biased violence, indicating a broadening in meaning and context.
Abuse (*abuso*)	12–13	1–2	0.00	The term is used in the context of *child abuse* in addition to its original associations with *power abuse*.
Abuse (*abuso*)	15–16	2–3	0.15	The term is used in the context of *sexual abuse* in addition to its original associations with *power abuse* and *child abuse*.
Climate (*clima*)	11–12	3–3	0.08	The term is mainly used for *environmental and climate issues* in addition to its previous usages for *a type of atmosphere* (e.g., political tension) or *a particular situation* (e.g., festive atmosphere).
Woman (*donna*)	8–9	2–3	0.28	In the 9th legislature, the term appears in relation to the bill for the establishment of voluntary military service for women in the *Italian Armed Forces.*
Gender (*genere*)	15–16	5–6	0.08	The term has evolved beyond its original usage as a means to denote a *kind* or *type* of something and has acquired a new connotation related to *gender identity* and *sexual gender*.
Seizure (*sequestro*)	5–6	1–2	0.03	The term underwent a semantic shift, expanding from its original meaning of *seizure* to also refer to the act of *person kidnapping*, due to the first kidnapping for extortion on December 18, 1972.

Table 5 Example of sense shift associated with the corresponding word, time interval, prominence and a short description

Word	Label	Time-interval	Prominence	Sense shift	Description
Clean (*pulito*)	Hygiene, purity, integrity	7–8	16–10	0.11	The sense nodule has undergone a "broadening" shift. In the 7th legislature, it was related to concepts like *honesty, moral correctness, fighting criminality*. In the 8th legislature its scope expanded to include *eliminating deception and pollution*, and *cleaning up the old regime*. In the 8th legislature, expressions like *clean sweep*, *clean country*, and *clean environment* emerge. This shift can be attributed to investigations such as "The Mani Pulite" and "Tangentopoli" scandals that revealed a fraudulent and corrupt system.
Environment (*ambiente*)	Environmental administration; environmental management; environmental protection	8–9	100–100	0.15	The sense nodule exhibited a"broadening" shift. In the 8th legislature, it was related to concepts like *political environment, work environment*. In the 9th legislature its scope expanded to include *ministerial issues* and *environmental bodies* for environmental protection. This shift can be attributed to the establishment of the Ministry of the Environment during the 9th legislature.
Right (*diritto*)	Law, human right; international right	7–8	26–33	0.17	The sense nodule exhibited a broadening shift. During the 7th legislature, it was primarily associated with concepts such as *law, legal norms*, and *human rights*. In the 8th legislature, its scope expanded specifically in relation to *human rights*. This shift can be attributed to the international agreement known as the Vienna Convention on the Law of Treaties. Indeed, expressions like *Vienna Convention* and *international law* emerged during the 7th legislature, while in the 8th legislature, expressions like *right of* emerged.

Table 5 (continued)

Word	Label	Time-interval	Prominence	Sense shift	Description
Party (*partito*)	Political parties; Left parties	11–12	96–97	0.11	The sense nodule exhibited a shift in meaning. During the 11th legislature, it was primarily associated with concepts such as *Left parties*, *political party*, and *transparency*. In the 12th legislature, its contextual scope expanded to include the idea of *coalition*. This shift can be attributed to the birth of the Italian People's Party. Terms like *Socialist Party* and *Democratic Party* emerged in the 8th legislature, while the 12th legislature witnessed the emergence of the expression *Italian People's Party*.
Violence (*violenza*)	Violence in social contexts	12–13	28–48	0.21	The sense nodules shifted, expanding from *physical violence* in the 12th legislature to also include *sexual assault* in the 13th legislature.
Opposition (*opposizione*)	Social opposition; political opposition	8–9	48–34	0.15	The sense nodule exhibited a narrowing shift in meaning. In the 8th legislature, it primarily pertained to the concept of *political opposition*. In the 9th legislature, its contextual expansion included a specific emphasis on *the role of political opposition* and *its significance as a critical* voice.
Abortion (*aborto*)	Numerical incidence and social implications of abortion	16–17	13–16	0.20	The sense nodule exhibited a narrowing shift, a shift in focus. In the 16th legislature, it was primarily associated with concepts such as *forced, illegal, and clandestine abortions*, as well as *women's healthcare*. During the 17th legislature, attention turned towards concern regarding the *rising number of medical staff who were conscientious objectors to providing* abortion and its potential impact on *increasing forced, illegal, and clandestine abortions*.

evaluation for RuShiftEval2021 was executed only on Subtask 2, since no binary labels are provided. Notably, the Russian corpus of RuShiftEval2021 spans three historical periods, allowing a further demonstration of WiDiD's effectiveness and robustness in detecting semantic shift over time. Note that no benchmarks are currently available over more than two multiple, consecutive time intervals.

Table 6 summarises the benchmarks considered.

6.1 Experimental setup

To evaluate WiDiD, we exploited the same setup described in Sect. 5.2 with the following modifications. We used a monolingual BERT model for each language, namely *bert-base-uncased* for English, *bert-base-italian-cased* for Italian, and *rubert-base-cased* for Russian. The models are base versions of BERT with 12 attention layers and 12 hidden layers of size 768. Furthermore, we compared the use of BERT models with two different multilingual models, both with 12 attention layers and 12 hidden layers of size 768, that is, mBERT *bert-base-multilingual-cased* and XLM-R *xlm-roberta-base*. As an exception, we only tested multilingual models for Latin since a monolingual model is not currently available.

Furthermore, going with the intuition that sense prototypes can be beneficial in limiting noise in the vector representations, we compared the use of JSD (described in Sect. 4) with the method based on sense nodules recently proposed by Kashleva et al. (2022). Following Kashleva et al. (2022), we define the semantic shift $\mathcal{S}_w$ as the average pairwise distance (APDP) between all pairs of the sense prototypes $\mu^t_{w,1..k} \in M^t_w$ and $\mu^{t-1}_{w,1..k} \in M^{t-1}_w$. Intuitively, the higher $\mathcal{S}_w$, the more the word w has shifted in meaning.

$$APDP(M^t_w, M^{t-1}_w) = \frac{\sum_{\mu^t_{w,i} \in M^t_w,\ \mu^{t-1}_{w,j} \in M^{t-1}_w} d(\mu^t_{w,i}, \mu^{t-1}_{w,j})}{|M^t_w||M^{t-1}_w|}$$

However, unlike (Kashleva et al. 2022), we set d as the Canberra distance instead of the cosine distance[9].

In line with previous work (Periti & Montanelli, 2024), for Subtask 1, we binarised the score of a word by using the threshold θ that maximises the overall result on the test set. Intuitively, the label 0 is assigned to a word if its JSD score is lower than θ, otherwise the label 1 is assigned to the word. It is worth noting that, development and training sets are not available for the majority of the benchmark, as SSD is typically framed in an unsupervised scenario (Schlechtweg et al., 2020). Therefore, the evaluation of Subtask 1 only provides an indication of the model's capability to recognize semantic shift. Indeed, the threshold is set based on the test set. This is also the reason why Subtask 2 is far more popular than Subtask 1. For Subtask 2, we directly used the JSD scores as degree of semantic shift.

[9] Empirical results in our experiments consistently demonstrated the superiority of using the Canberra distance over the Cosine Distance.

Table 6 Period, size in tokens, reference, and number of target words for the evaluation benchmark considered

SemEval		Periods	Tokens	References	Target words
English	$C_1 C_2$	1810–1860 1960–2010	6 M 6 M	Schlechtweg et al. (2020)	37
Latin	$C_1 C_2$	– 200–0 0–2000	65 k 253 k	Schlechtweg et al. (2020)	40
German	$C_1 C_2$	1800–1899 1946–1990	70.2 M 72.3 M	Schlechtweg et al. (2020)	48
Swedish	$C_1 C_2$	1790–1830 1895–1903	71.0 M 110.0 M	Schlechtweg et al. (2020)	31
DIACRIta					
Italian	$C_1 C_2$	1945–1970 1990–2014	52 M 196 M	Basile et al. (2020)	18
RuShiftEval					
Russian	$C_1 C_2 C_3$	1700–1916 1918–1990 1992–2016	94 M 123 M 107 M	Kutuzov and Pivovarova (2021)	99
LSCDiscovery					
Spanish	$C_1 C_2$	1810–1906 1994–2020	13.0 M 22.0 M	Zamora-Reina et al. (2022)	100

6.2 Experimental results

For the sake of comparison, we report the top state-of-the-art results achieved using contextualised embeddings for Subtask 1 and Subtask 2 in Tables 7 and 8, respectively. To ensure a fair comparison, we exclusively report results obtained by unsupervised approaches leveraging contextualised embeddings. In addition, it is worth noting that we are reporting the best result achieved in multiple experiments (e.g., using different models and measures). Accordingly, we have compared our best results with the provided state-of-the-art results.

Table 9 presents the results of our evaluation for both Subtask 1 and 2.

For Subtask 1, we note that our results have the potential to outperform the results shown in Table 7 across all evaluated benchmarks. Specifically, for the DIACRIta benchmark, which is relevant for our study due to the shared language of our case study corpus, both BERT+JSD and mBERT+JSD exhibit equal effectiveness by correctly labelling 17 out of 18 words.

For Subtask 2, our results outperform state-of-the-art results for English and Russian, while being comparable with the state-of-the-art results for the other benchmarks.

As a general remark, and in line with the finding of Kutuzov and Giulianelli (2020), we note that the measure which produces a more uniform predicted score distribution (APDP) works better for the test sets with skewed gold distributions, and the measure which produces a more skewed predicted score distribution (JSD) works better for the uniformly distributed test sets.

As for the model comparison, we observed that, on average, different models achieve similar results for Subtask 1. However, the selection of the model is crucial for Subtask 2. For instance, both BERT and XLM-R demonstrate good performance for English, while the use of mBERT leads to significantly worse results. Interestingly, contrary to the widespread belief that monolingual models are more suitable than multilingual ones, we found that only for English (Subtask 2) and Spanish (Subtask 1 and 2) did employing a monolingual BERT model prove more effective than using a multilingual model. Additionally, despite the expectation that XLM-R would outperform mBERT due to the larger amount of training data and parameters it uses, we observed that mBERT is the most suitable model for Latin (Subtask 1) and Russian (Subtask 2).

7 Discussion and conclusion

7.1 Data quality

One crucial aspect of diachronic corpora is that the number of documents is often imbalanced, and the presence of a target word is not equally reflected in all the time points considered. In common scenarios, more documents are available for more recent time periods and *it may not be possible to achieve balance in the sense expected from a modern corpus* (Tahmasebi & Dubossarsky, 2023). Furthermore,

Table 7 Subtask 1: accuracy scores achieved from various state-of-the-art experiments

References	SemEval				DiacrIta
	English $C1 - C2$	*Latin* $C1 - C2$	*German* $C1 - C2$	*Swedish* $C1 - C2$	*Italian* $C1 - C2$
Unsupervised					
Kanjirangat et al. (2020)	.541	.375	.708	.742	–
Martinc et al. (2020)	.703*	.700	.667*	.710*	–
Karnysheva and Schwarz (2020)	.568	.650	.583	.645	–
Rother et al. (2020)	.622	.575	.729	.742	–
Cuba Gyllensten et al. (2020)	.568	.675	.562	.710	–
Wang et al. (2020)	–	–	–	–	.610*
Giulianelli et al. (2022)	.459*	.500*	.521*	– .516*	.389*
Supervised					
Ma et al. (2024)	.784	.700	.813	.806	–
WiDiD	**.757**	**.750**	**.729**	**.774**	**.944**

Asterisks denote scores obtained via fine-tuning contextualised models, while hyphens indicate unavailable experimental results. Bold denotes the best unsupervised scores

Table 8 Subtask 2: Spearman's correlation coefficients achieved from various state-of-the-art experiments

References	SemEval				LSCDiscovery	RuShiftEval		
	English $C1-C2$	*Latin* $C1-C2$	*German* $C1-C2$	*Swedish* $C1-C2$	*Spanish* $C1-C2$	*Russian* $C1-C2$	*Russian* $C2-C3$	*Russian* $C1$-$C3$
Unsupervised								
Kanjirangat et al. (2020)	.159	.231	.525	.141	–	–	–	–
Martinc et al. (2020)	.436*	.481	.528*	.238*	–	–	–	–
Karnysheva and Schwarz (2020)	.155	.177	.388	.062	–	–	–	–
Rother et al. (2020)	.306	.321	.605	.268	–	–	–	–
Cuba Gyllensten et al. (2020)	.209	.399	**.656**	.234	–	–	–	–
Montariol et al. (2021)	.456*	**.488***	.561*	**.561***	–	–	–	–
Giulianelli et al. (2022)	.127*	.318*	.287*	− .108*	–	.247*	.267*	.362*
Kashleva et al. (2022)	–	–	–	–	**.553***	–	–	–
Supervised								
Aida and Bollegala (2024)	.774	.124	.902	.656	–	.805	.811	.846
Cassotti et al. (2023)	.757	− .056	.877	.754	–	.799	.833	.842
WiDiD	**.651**	.433	.527	.499	.544	**.273**	**.393**	**.407**

Asterisks denote scores obtained via fine-tuning contextualised models, while hyphens indicate unavailable experimental results. Bold denotes the best unsupervised scores

the quality of the analysed data can significantly influence the results. Similar to the imbalance issue, the quality of the data is generally higher for recent documents than for past documents. Old documents are often digitised as images using an OCR scanning process to convert them into text. However, this procedure can introduce *OCR errors* that contribute to degrading the quality of the analysis.

In our case study corpus, the imbalance was caused by the inherent varying duration of legislatures rather than the availability of documents. A legislature is usually associated with a time period of up to 5 years, which corresponds to the duration of an election cycle. However, in cases where the Parliament withdraws its support from the government through a *vote of no confidence*, the duration can be shorter.

In terms of data quality, the documents in our case study corpus were originally stored as images and digitised through an OCR scanning process. As a result, several characters were misrecognised, omitted, or erroneously inserted, distorting the original text across all the legislatures. Although a precise estimation of the extent of these errors is currently unavailable, we enforced heuristics to mitigate OCR errors and retain only the highest-quality sentences in the corpus. Despite the efforts to remove highly corrupted sentences, some errors persist and the processing has further increased the existing imbalance in the corpus.

These issues affect the quality of contextualised embeddings generated by BERT-like models. Thus far, only a few studies have explored the influence of OCR errors on contextualised embeddings (Todorov & Colavizza, 2022; Jiang et al., 2021). As a result, the impact of OCR errors on contextualisation remains unclear, and quantifying their effect is challenging. Nevertheless, we hypothesise that there might be significant side effects. For instance, one common problem caused by OCR errors is the inconsistent use of punctuation, resulting in longer or shorter sentences that degrade the quality of the embeddings. Additionally, OCR often introduces or removes spaces, which disrupts sentence segmentation. For example, the word “aperitivo" (happy hour) may become a three-word expression like “ape re timo” (in English, *bee king thyme*), thus affecting the correct interpretation of the sentence. The meaning of words can be also altered by OCR errors that remove accents. For instance, “papa" and “papà" have different meanings (*pope* and *father*, respectively).

In a study on diachronic word sense discrimination, Tahmasebi et al. (2013) showed that due to the design of the algorithm, the quality of the clusters did not degrade with decreasing quality of the corpus, but the number of clusters was radically reduced. When using contextualised embeddings this is not the case, since we can produce embeddings for each occurrence of a target word regardless of the quality of the sentence. As long as the word we are interested in is correctly spelled, its contextual representation will contribute to the meaning of the word, however, with reduced quality. Thus, with contextualised embeddings, the quality of the output inherently depends on the quality of the input data. Due to the significant number of OCR errors in our case study, our empirical results may be less accurate and reliable. However, we expect the OCR errors to affect the corpus at each time period roughly evenly, and thus all senses of a word should be affected to the same degree in any given time period. As a result, small clusters may not be detected and some clusters could show up later than expected. Nevertheless, the case study serves its purpose in demonstrating how WiDiD works in a concrete application, and **is not**

Table 9 Evaluation scores for Subtask 1 and Subtask 2 achieved via accuracy (Acc) and Spearman's correlation coefficients (Corr), respectively, over different benchmarks and setups

JSD/APDP	SemEval				LSCDiscovery	RuShiftEval			DiacrIta
	*English C*1 − *C*2	*Latin C*1 − *C*2	*German C*1 − *C*2	*Swedish C*1 − *C*2	*Spanish C*1 − *C*2	*Russian C*1 − *C*2	*Russian C*2 − *C*3	*Russian C*1 − *C*3	*Italian C*1 − *C*2
Acc Sub. 1									
BERT	.622/.730	.675/.625	**.729**/.708	.742/**.774**	**.688**/**.688**	–	–	–	**.944**/.833
mBERT	.649/.676	**.750**/.675	**.729**/.646	.742/**.774**	.675/.638	–	–	–	**.944**/.722
XLM-R	.622/**.757**	.725/.650	**.729**/.708	**.774**/**.774**	.675/.625	–	–	–	.889/.833
Corr Sub. 2									
BERT	.256/**.651**	.334/.165	.407/.363	.012/.155	.429/**.544**	.198/.204	.265/.238	.271/.177	–
mBERT	.244/.237	.410/-.093	.397/.280	.015/.132	.450/.420	.263/**.273**	.348/**.393**	.398/**.407**	–
XLM-R	.291/.635	**.433**/-.096	.225/**.527**	.087/**.499**	.463/.322	.021/.132	.328/.250	.292/.256	–

For each benchmark, we report our results obtained by using different contextualised models (i.e, BERT, mBERT, XLM-R) and different semantic shift measures (i.e., **JSD/APDP**). We report in bold the highest scores for each benchmark and subtask

meant as an in-depth, exploratory social and linguistics study of the Italian parliament. In the following, we outline some limitations related to our case study. Specifically, we pre-defined the set of target words to consider in the analysis, without applying WiDiD to the entire vocabulary. Moreover, since the case study focuses on semantic change in a specific context (i.e., the Italian Parliament), the meanings of the target words occurring in the corpus could be somehow limited. Finally, the use of pretrained language models like BERT can represent a limitation: such models are typically trained on corpora that differ significantly in terms of topics and time periods from the domain under consideration.

7.2 Incremental semantic shift detection

Incremental semantic shift detection enables a more fine-grained analysis of semantic shift by tracing the evolution of different word meanings over time. However, semantic shift is not uniform across all words or domains. Some words may experience rapid shift in meaning, while others can change gradually or remain relatively stable. Therefore, computational approaches need to be flexible enough to handle both short- and long-term semantic shift. In addition, word meanings do not necessarily change in a linear way. They are not strictly limited to increasing, decreasing, or remaining stable in prominence. Instead, word meanings can be influenced by various circumstances, leading to both regular and irregular trends that can activate or deactivate meanings in different time periods. These properties make a complete modelling of semantic shift extremely complex. While we are advancing existing state-of-the-art change detection methods significantly, we have reduced the complexity in several ways and made several design choices that can affect the results. We discuss a few of these choices below.

First, we chose not to perform online clustering of elements (i.e., sentences with a target word) one-by-one but instead to consider all elements stemming from a time period at the same time. Conducting the clustering step of WiDiD after adding a single new element would enforce clustering on a small number of elements, namely the newly added element and the previous n sense prototypes. Such a procedure, that does not correspond to our typical research scenario, is unlikely to result in converging clusters and can lead to erroneously merged clusters, thus losing the"memory" already gathered. We thus opted to cluster all elements from a time period together with the previous sense prototypes all at once, leading to more robust clustering results. While this procedure increases the overall amount of data during clustering, it does not handle gradual semantic change, where only a few elements of a new cluster may initially be present. Consequently, recognition of a semantic shift is likely to occur at a later stage, when a consistent amount of evidence supporting the change is considered. Specifically, if the evidence for capturing a new sense (i.e., creating a new sense cluster) is insufficient within a specific time period, WiDiD will misclassify such evidence. However, as a feature of the *what is done is done* approach, an assignment will be never reconsidered even if additional evidence becomes available in later time periods. As a consequence, in order to recognize a new

sense, a substantial evidence of that sense must appear in a specific time period, rather than a cumulative evidence across all the processed periods. A similar issue may occur when the evidence for capturing a new sense is sufficient in a certain time period, but some word occurrences denoting the new sense are incorrectly associated by WiDiD with other active senses. This misclassification can lead to a downsample of evidence for the new sense, causing it to be underrepresented and not recognised until more supporting evidence becomes available in later time periods. The characteristics of the data under analysis must guide the iteration frequency of WiDiD over time to reduce disambiguation errors and minimise the overlooking of emerging senses. To overcome this issue, the combination of WiDiD with a global evolutionary clustering approach can be enforced to introduce the possibility to review past assignments and reverse them if needed.

In WiDiD each sense nodule is currently represented by a single-sense prototype representation, with the same importance as a new element (i.e., contextualised embedding of a word). This approach leads to a higher risk of sense nodules being merged or confused over time. Empirical results indicate that while some clusters persist over time even without the integration of new elements, the majority tend to merge with other clusters over time. In the final step this results in an increase in the number of clusters stemming from the last time period and a decrease in the number of clusters stemming from earlier periods (since in the earlier time periods there were more opportunities for merging). While the aggregation of sense nodules may sometimes aid in focusing on lexicographic meaning (rather than just on sense nodules), at other times it results only in noise representations. This problem could possibly be solved by using a different weighting schema for sense nodules and new elements, but manually annotated ground truth data is needed to perform large-scale evaluation so as to choose the best weighting schema.

In the current implementation, WiDiD considers all the occurrences of a word and preserves all the generated sense nodules (i.e., clusters). A pre-processing step can be introduced at the beginning of the WiDiD approach to discard *ambiguous* word occurrences due to OCR errors and/or limited contexts that prevents to capture the appropriate meaning of the word. Similarly, a post-processing step can be applied to refine the memory of active meanings at the end of each Incremental Clustering step. For instance, post-processing can be used over cluster integrations to distinguish between *valid updates* (e.g., active clusters enriched with at least n elements), and *invalid updates* (e.g., active clusters enriched with fewer than n elements). Post-processing can also be employed in cluster merging to decide when it is appropriate to consolidate two or more sense clusters into a single one. Yet, post-processing can be employed to classify sense clusters as "lost" or no longer active (that can be forgotten). For example, each cluster can be associated with an *aging index* to measure how recently it has been updated and to decide when it should be considered lost and removed from memory (Castano et al., 2024; Periti et al., 2022). In general, both pre- and post-processing steps can be managed through the use of thresholds whose value must be customized according to the considered dataset (e.g., size, domain, time periods, style), and the nature of semantic change under analysis. As a matter of example, in case studies with limited or high-quality data, a cluster integration of one or a few elements might be a valid update; whereas

in studies with extensive or medium-quality data, such minor updates could be considered noisy and thus ignored. Similarly, in case studies where the focus is on detecting immediate meaning changes, such as in rapidly evolving fields, a few intervals without cluster integrations may be sufficient to deem a sense cluster as lost; conversely, when the focus is on periodic senses (e.g., the meaning of *gold* during Olympics games), prematurely pruning senses from the memory could lead to capture as changed a meaning that is only appearing and disappearing from memory (Periti & Tahmasebi, 2024).

When it comes to interpreting semantic shift across multiple time points, two different approaches can be adopted: a posteriori analysis and evolutionary analysis. In a posteriori analysis, the snapshot associated with the clustering result of the last iteration is used. Thus, the cluster membership distribution across different time points is considered with respect to the clustering result of the final iteration. That is, we do not consider two clusters individually in previous time periods if they have been merged by the last time period. This analysis focuses on examining how the clusters are distributed and assigned across time, providing insights into the temporal patterns of semantic shift and is a simplification of the full semantic shift problem. Evolutionary analysis, on the other hand, emphasises the behaviour of the clusters themselves rather than their specific distribution across time. It investigates the evolution of clusters, such as their merging or integration over time. Observing changes in cluster composition and structure can yield valuable information regarding the dynamic nature of semantic shift (Hu et al., 2019).

In our specific case study, we used a posteriori analysis and chose not to apply any threshold mechanism, as it was convenient for illustrating the applicability of WiDiD to our case study and the complete history of each cluster during the considered time periods. We are currently working on developing techniques to present the patterns captured by *evolutionary analysis* (i.e., incremental analysis of new sense nodules, their merging and integration). However, such analysis requires large-scale evaluation across multiple time points and is significantly more complex. To be a useful research tool, evolutionary analysis also requires ways to represent the results without overloading the user. We are currently working on creating evaluation data for such a scenario.

Finally, recent research has demonstrated that embeddings lie in an anisotropic space, indicating that all vectors are within a narrow cone. The consequence is that even embeddings of unrelated words are close together in distributional space and thus exhibit very high similarity. As a result, if a sense prototype is even slightly distorted, one or more sense prototypes may be incorrectly clustered and the algorithm's results may exhibit a large degree of randomness. A way to overcome this issue might be to project the embeddings onto a larger part of the space (i.e., making the cone wider), thus creating more distance between elements.

7.3 Possible applications of WiDiD

Both historical linguistics and lexicography involve the direct application of semantic shift detection. The former compares change patterns across time and languages,

and the latter needs to update dictionary entries on the basis of new information from modern or historical texts. Much of this work requires manually labelling and interpreting each cluster, which can be a time-consuming task, especially when there are large sets of clusters or when many words are considered at once.

We envision a Query Answering system based on WiDiD as a solution to facilitate the interpretation of semantic shift and the analysis of specific word meanings over time. WiDiD allows for intelligent filtering, both on the word level and the sense level. For example, one could study particular words in certain periods of time (pre- and post-war, or pre- and post-pandemic are typical periods of study). Alternatively, one could investigate all documents that use a word in a specific sense.

Such fine-grained analysis across temporal dimensions and all senses of a word is an extremely useful tool in research fields where diachronic analysis of word meaning is central. It is, however, important to couple the outcome of an approach like WiDiD with confidence values that reflect the level of certainty associated with an unsupervised model trained on text of varying quality.

7.4 Concluding remarks

In this paper, we extend a recent approach to Semantic Shift Detection called WiDiD (Periti et al., 2022) by (i) adding cluster analysis techniques and visualisation; (ii) providing a practical demonstration of our extended approach; (iii) conducting a comprehensive evaluation; and (iv) engaging in detailed discussion of the WiDiD usage.

We employ the WiDiD algorithm because it is the first incremental and scalable approach based on evolutionary clustering of contextualised word embeddings to model the evolution of word meaning over time and detect lexical semantic change. We demonstrate the practical application of WiDiD on a diachronic corpus of Italian parliamentary speeches spanning eighteen distinct time periods. We evaluated the performance of WiDiD over seven popular labeled benchmarks. Our empirical results show that, for certain languages, WiDiD outperforms state-of-the-art approaches, while achieving comparable results for other languages. At the same time, WiDiD captures significantly more information, thus allowing more in-depth analysis of the detected change than existing approaches to semantic shift detection. We believe this paper holds significant relevance in **changing the course of the current modeling of semantic shift**, where, currently, the temporal nature of the documents is generally disregarded. With this paper, we aim to pave the way for further work that relies on incremental/evolutionary clustering algorithms to model lexical semantic change by considering the temporal nature of the documents under consideration.

Acknowledgements This work has in part been funded by the project Towards Computational Lexical Semantic Change Detection supported by the Swedish Research Council (2019-2022; contract 2018-01184), and in part by the research program Change is Key! supported by Riksbankens Jubileumsfond (under reference number M21-0021). The computational resources were provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS).

Author contributions *Francesco Periti*—conceptualization; methodology; software; validation; investigation; writing—original draft; writing—review and editing; *Sergio Picascia*—software; visualization; data curation; *Stefano Montanelli*—supervision; methodology; data curation; *Alfio Ferrara*—methodology; data curation; *Nina Tahmasebi*—supervision; writing—original draft; writing—review and editing; funding acquisition.

Funding Open access funding provided by Università degli Studi di Milano within the CRUI-CARE Agreement. This work has in part been funded by the project Towards Computational Lexical Semantic Change Detection supported by the Swedish Research Council (2019-2022; contract 2018-01184), and in part by the research program Change is Key! supported by Riksbankens Jubileumsfond (under reference number M21-0021). The computational resources were provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS).

Data availibility Italian parliamentary speeches have been downloaded from https://data.camera.it/.

Code availability Code is available at https://github.com/FrancescoPeriti/WiDiD.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

Consent for publication The authors gave explicit consent to submit.

Ethical approval and consent to participate The authors have nothing to disclose.

References

Aida, T., Bollegala, D.: A Semantic Distance Metric Learning approach for Lexical Semantic Change Detection. In: Ku, L.-W., Martins, A., Srikumar, V. (eds.) Findings of the Association for Computational Linguistics ACL 2024, pp. 7570–7584. Association for Computational Linguistics, Bangkok, Thailand and virtual meeting (2024). https://aclanthology.org/2024.findings-acl.451

Alkhalifa, R., Kochkina, E., & Zubiaga, A. (2023). Building for tomorrow: Assessing the temporal persistence of text classifiers. *Information Processing & Management, 60*(2), 103200. https://doi.org/10.1016/j.ipm.2022.103200

Azarbonyad, H., Dehghani, M., Beelen, K., Arkut, A., Marx, M., & Kamps, J. (2017). Words are malleable: Computing semantic shifts in political and media discourse. In *Proceedings of the 2017 ACM on conference on information and knowledge management* (*CIKM '17*) (pp. 1509–1518). Association for Computing Machinery. https://doi.org/10.1145/3132847.3132878

Basile, P., Caputo, A., Caselli, T., Cassotti, P., & Varvara, R. (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*). CEUR-WS.org. https://ceur-ws.org/Vol-2765/paper158.pdf

Basile, P., Caputo, A., Luisi, R., & Semeraro, G. (2016). Diachronic analysis of the Italian language exploiting Google Ngram. In A. Corazza, S. Montemagni, & G. Semeraro (Eds.), *Proceedings of*

the third Italian conference on computational linguistics CLiC-It 2016. Accademia University Press. Digital reference of the book. https://doi.org/10.4000/books.aaccademia.1707

Basile, P., Semeraro, G., & Caputo, A. (2019). Kronos-it: A dataset for the Italian semantic change detection task. In *CLiC-it*. CEUR-WS.org. https://ceur-ws.org/Vol-2481/paper3.pdf

Bloomfield, L. (1933). *Language*. Holt, Rinehart and Winston

Boros, E., Ehrmann, M., Romanello, M., Najem-Meyer, S., & Kaplan, F. (2024). Post-correction of historical text transcripts with large language models: An exploratory study. In Y. Bizzoni, S. Degaetano-Ortlieb, A. Kazantseva, & S. Szpakowicz (Eds.), *Proceedings of the 8th joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature* (*LaTeCH-CLfL 2024*) (pp. 133–159). Association for Computational Linguistics. https://aclanthology.org/2024.latechclfl-1.14

Card, D. (2023). Substitution-based semantic change detection using contextual embeddings. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics* (Vol 2: Short Papers, pp. 590–602). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-short.52 . https://aclanthology.org/2023.acl-short.52

Cassotti, P., Siciliani, L., DeGemmis, M., Semeraro, G., & Basile, P. (2023). XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In *Proceedings of the 61st annual meeting of the association for computational linguistics* (Vol. 2: Short Papers, pp. 1577–1585). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-short.135

Castano, S., Ferrara, A., Montanelli, S., & Periti, F. (2024). Incremental affinity propagation based on cluster consolidation and stratification. https://doi.org/10.48550/arXiv.2401.14439

Cruse, D. A. (2000). *Aspects of the micro-structure of word meanings*. Oxford University Press.

Cuba Gyllensten, A., Gogoulou, E., Ekgren, A., & Sahlgren, M. (2020). SenseCluster at SemEval-2020 Task 1: Unsupervised lexical semantic change detection. In A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, & E. Shutova (Eds.), *Proceedings of the fourteenth workshop on semantic evaluation* (pp. 112–118). International Committee for Computational Linguistics. https://doi.org/10.18653/v1/2020.semeval-1.12

Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (*EMNLP-IJCNLP*) (pp. 55–65). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1006 . https://aclanthology.org/D19-1006

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science, 315*(5814), 972–976. https://doi.org/10.1126/science.1136800

Geeraerts, D. (2020). Semantic change: "What the Smurf?". The Wiley Blackwell companion to semantics (pp. 1–24). https://doi.org/10.1002/9781118788516.sem042

Giulianelli, M., Del Tredici, M., & Fernández, R. (2020). Analysing lexical semantic change with contextualised word representations. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 3960–3973). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.365

Giulianelli, M., Kutuzov, A., & Pivovarova, L. (2022). Do not fire the linguist: Grammatical profiles help language models detect semantic change. In N. Tahmasebi, S. Montariol, A. Kutuzov, S. Hengchen, H. Dubossarsky, & L. Borin (Eds.), *Proceedings of the 3rd workshop on computational approaches to historical language change* (pp. 54–67). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.lchange-1.6

Hu, R., Li, S., & Liang, S. (2019). Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 3899–3908). Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/P19-1379

Jiang, M., Hu, Y., Worthey, G., Dubnicek, R. C., Underwood, T., & Downie, J. S. (2021). Impact of OCR quality on BERT embeddings in the domain classification of book excerpts. In *Proceedings of the conference on computational humanities research* 2021, Amsterdam, the Netherlands. https://ceur-ws.org/Vol-2989/long_paper43.pdf

Kanjirangat, V., Mitrovic, S., Antonucci, A., & Rinaldi, F. (2020). SST-BERT at SemEval-2020 Task 1: Semantic shift tracing by clustering in BERT-based embedding spaces. In A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, & E. Shutova (Eds.), *Proceedings of the fourteenth workshop on*

semantic evaluation (pp. 214–221). International Committee for Computational Linguistics. https://doi.org/10.18653/v1/2020.semeval-1.26 . https://aclanthology.org/2020.semeval-1.26

Karnysheva, A., & Schwarz, P. (2020). TUE at SemEval-2020 task 1: Detecting semantic change by clustering contextual word embeddings. In A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, & E. Shutova (Eds.) *Proceedings of the fourteenth workshop on semantic evaluation* (pp. 232–238). International Committee for Computational Linguistics. https://doi.org/10.18653/v1/2020.semeval-1.28 . https://aclanthology.org/2020.semeval-1.28

Kashleva, K., Shein, A., Tukhtina, E., & Vydrina, S. (2022). HSE at LSCDiscovery in Spanish: Clustering and profiling for lexical semantic change discovery. In N. Tahmasebi, S. Montariol, A. Kutuzov, S. Hengchen, H. Dubossarsky, & L. Borin (Eds.), *Proceedings of the 3rd workshop on computational approaches to historical language change* (pp. 193–197). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.lchange-1.21 . https://aclanthology.org/2022.lchange-1.21

Kazi, R., Amato, A., Wang, S., & Bucur, D. (2022). Visualisation methods for diachronic semantic shift. In A. Cohan, G. Feigenblat, D. Freitag, T. Ghosal, D. Herrmannova, P. Knoth, K. Lo, P. Mayr, M. Shmueli-Scheuer, A. Waard, & L.L. Wang (Eds.), *Proceedings of the third workshop on scholarly document processing* (pp. 89–94). Association for Computational Linguistics. https://aclanthology.org/2022.sdp-1.10

Kutuzov, A., & Giulianelli, M. (2020). UiO-UvA at SemEval-2020 Task 1: Contextualised embeddings for lexical semantic change detection. In A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, & E. Shutova (Eds.), *Proceedings of the fourteenth workshop on semantic evaluation* (pp. 126–134). International Committee for Computational Linguistics. https://doi.org/10.18653/v1/2020.semeval-1.14

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E.(2018). Diachronic word embeddings and semantic shifts: A survey. In E.M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th international conference on computational linguistics* (pp. 1384–1397). Association for Computational Linguistics. https://aclanthology.org/C18-1117

Kutuzov, A., & Pivovarova, L. (2021). RuShiftEval: A shared task on semantic shift detection for Russian. In *Proceedings of the conference on computational linguistics and intellectual technologies (dialogue)*. RSUH. https://www.dialog-21.ru/media/5536/pivovarovalpluskutuzova151.pdf

Kutuzov, A., Pivovarova, L., & Giulianelli, M. (2021). Grammatical profiling for semantic change detection. In A. Bisazza & O. Abend (Eds.) *Proceedings of the 25th conference on computational natural language learning* (pp. 423–434). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.conll-1.33 . https://aclanthology.org/2021.conll-1.33

Kutuzov, A., Velldal, E., & Øvrelid, L. (2022). Contextualized embeddings for semantic change detection: Lessons learned. In L. Derczynski (Ed.), *Northern European Journal of Language Technology* (Vol. 8). Northern European Association of Language Technology. https://doi.org/10.3384/nejlt.2000-1533.2022.3478 . https://aclanthology.org/2022.nejlt-1.9

Ma, X., Strube, M., Zhao, W.: Graph-based Clustering for Detecting Semantic Change Across Time and Languages. In: Graham, Y., Purver, M. (eds.) Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1542–1561. Association for Computational Linguistics, St. Julian's, Malta (2024). https://aclanthology.org/2024.eacl-long.93

Martinc, M., Montariol, S., Zosa, E., & Pivovarova, L. (2020). Capturing evolution in word usage: Just add more clusters? In *Companion proceedings of the web conference 2020. WWW '20* (pp. 343–349). Association for Computing Machinery. https://doi.org/10.1145/3366424.3382186 .

Martinc, M., Montariol, S., Zosa, E., & Pivovarova, L. (2020). Discovery team at SemEval-2020 task 1: Context-sensitive embeddings not always better than static for semantic change detection. In A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, & E. Shutova (Eds.), *Proceedings of the fourteenth workshop on semantic evaluation* (pp. 67–73). International Committee for Computational Linguistics. https://doi.org/10.18653/v1/2020.semeval-1.6 . https://aclanthology.org/2020.semeval-1.6

Martinc, M., Novak, P. K., & Pollak, S. (2020). Leveraging contextual embeddings for detecting diachronic semantic shift. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the twelfth language resources and evaluation conference* (pp. 4811–4819). European Language Resources Association. https://aclanthology.org/2020.lrec-1.592

Montariol, S., Martinc, M., & Pivovarova, L. (2021). Scalable and interpretable semantic change detection. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, & Y. Zhou (Eds.) *Proceedings of the 2021 conference of the North*

American Chapter of the Association for Computational Linguistics: Human language technologies (pp. 4642–4652). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.369 . https://aclanthology.org/2021.naacl-main.369

Noble, B., Sayeed, A., Fernández, R., & Larsson, S. (2021). Semantic shift in social networks. In L.-W. Ku, V. Nastase, & I. Vulić (Eds.), *Proceedings of *SEM 2021: The tenth joint conference on lexical and computational semantics* (pp. 26–37). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.starsem-1.3 . https://aclanthology.org/2021.starsem-1.3

Periti, F., Cassotti, P., Dubossarsky, H., Tahmasebi, N.: Analyzing Semantic Change through Lexical Replacements. In: Ku, L.-W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 4495–4510. Association for Computational Linguistics, Bangkok, Thailand (2024). https://aclanthology.org/2024.acl-long.246

Periti, F., Ferrara, A., Montanelli, S., & Ruskov, M. (2022). What is Done is Done: an Incremental Approach to Semantic Shift Detection. In N. Tahmasebi, S. Montariol, A. Kutuzov, S. Hengchen, H. Dubossarsky, & L. Borin (Eds.), *Proceedings of the 3rd workshop on computational approaches to historical language change* (pp. 33–43). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.lchange-1.4 . https://aclanthology.org/2022.lchange-1.4

Periti, F., & Montanelli, S. (2024). Lexical semantic change through large language models: A survey. *ACM Computing Surveys, 56*(11). https://doi.org/10.1145/3672393

Periti, F., & Tahmasebi, N. (2024). Towards a complete solution to lexical semantic change: An extension to multiple time periods and diachronic word sense induction. In N. Tahmasebi, S. Montariol, A. Kutuzov, D. Alfter, F. Periti, P. Cassotti, & N. Huebscher (Eds.), *Proceedings of the 5th workshop on computational approaches to historical language change* (pp. 108–119). Bangkok, Thailand: Association for Computational Linguistics. https://aclanthology.org/2024.lchange-1.10

Rachinskiy, M., & Arefyev, N. (2022). GlossReader at LSCDiscovery: Train to select a proper gloss in English—Discover lexical semantic change in Spanish. In N. Tahmasebi, A. Kutuzov, S. Hengchen, H. Dubossarsky, & L. Borin (Eds.), *Proceedings of the 3rd workshop on computational approaches to historical language change* (pp. 198–203). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.lchange-1.22. https://aclanthology.org/2022.lchange-1.22

Rodina, J., Trofimova, Y., Kutuzov, A., & Artemova, E. (2021). ELMo and BERT in semantic change detection for Russian. In W. M. P. Aalst, V. Batagelj, D. I. Ignatov, M. Khachay, O. Koltsova, A. Kutuzov, S. O. Kuznetsov, I. A. Lomazova, N. Loukachevitch, A. Napoli, A. Panchenko, P. M. Pardalos, M. Pelillo, A. V. Savchenko, & E. Tutubalina (Eds.), *Analysis of images, social networks and texts* (pp. 175–186). Springer.

Rother, D., Haider, T., & Eger, S. (2020). CMCE at SemEval-2020 Task 1: Clustering on manifolds of contextualized embeddings to detect historical meaning shifts. In A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, & E. Shutova (Eds.), *Proceedings of the fourteenth workshop on semantic evaluation* (pp. 187–193). International Committee for Computational Linguistics. https://doi.org/10.18653/v1/2020.semeval-1.22 . https://aclanthology.org/2020.semeval-1.22

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. (2020). SemEval-2020 task 1: Unsupervised lexical semantic change detection. In A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, & E. Shutova, (Eds.), *Proceedings of the fourteenth workshop on semantic evaluation* (pp. 1–23). International Committee for Computational Linguistics. https://doi.org/10.18653/v1/2020.semeval-1.1 . https://aclanthology.org/2020.semeval-1.1

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. In K. Erk, & N. A. Smith (Eds.), *Proceedings of the 54th annual meeting of the association for computational linguistics* (Vol. 1: Long Papers, pp. 1715–1725). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1162 . https://aclanthology.org/P16-1162

Su, Z., Tang, Z., Guan, X., Wu, L., Zhang, M., & Li, J. (2022). Improving temporal generalization of pre-trained language models with lexical semantic change. In Goldberg, Y., Kozareva, Z., & Zhang, Y. (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 6380–6393). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates. https://doi.org/10.18653/v1/2022.emnlp-main.428 . https://aclanthology.org/2022.emnlp-main.428

Tahmasebi, N., Borin, L., & Jatowt, A. (2021). Survey of computational approaches to lexical semantic change detection. Language Science Press. https://doi.org/10.5281/zenodo.5040302 .

Tahmasebi, N., & Dubossarsky, H. (2023). Computational modeling of semantic change. https://doi.org/10.48550/arXiv.2304.06337 . https://arxiv.org/abs/2304.06337

Tahmasebi, N., Niklas, K., Zenz, G., & Risse, T. (2013). On the applicability of word sense discrimination on 201 years of modern English. *International Journal on Digital Libraries, 13*(3–4), 135–153. https://doi.org/10.1007/s00799-013-0105-8

Tang, X. (2018). A state-of-the-art of semantic change computation. *Natural Language Engineering, 24*(5), 649–676. https://doi.org/10.1017/S1351324918000220

Todorov, K., & Colavizza, G. (2022). An assessment of the impact of OCR noise on language models. https://doi.org/10.48550/arXiv.2202.00470 . https://arxiv.org/abs/2202.00470

Wang, B., Di Buccio, E., & Melucci, M. (2020). University of Padova @ DIACR-Ita. In *Proceedings of the seventh evaluation campaign of natural language processing and speech tools for Italian. Final workshop* (*EVALITA 2020*). CEUR-WS, Marrakech, Morocco. https://ceur-ws.org/Vol-2765/paper91.pdf

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., & Rush, A. Transformers: State-of-the-art natural language processing. In Q. Liu, & D. Schlangen (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 38–45). Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6 . https://aclanthology.org/2020.emnlp-demos.6

Zamora-Reina, F. D., Bravo-Marquez, F., & Schlechtweg, D. (2022). LSCDiscovery: A shared task on semantic change discovery and detection in Spanish. In N. Tahmasebi, S. Montariol, A. Kutuzov, S. Hengchen, H. Dubossarsky, & L. Borin (Eds.), *Proceedings of the 3rd workshop on computational approaches to historical language change* (pp. 149–164). Association for Computational Linguistics, Dublin, Ireland (2022). https://doi.org/10.18653/v1/2022.lchange-1.16 . https://aclanthology.org/2022.lchange-1.16

Authors and Affiliations

Francesco Periti[1] · Sergio Picascia[1] · Stefano Montanelli[1] · Alfio Ferrara[1] · Nina Tahmasebi[2]

✉ Francesco Periti
francesco.periti@unimi.it

Sergio Picascia
sergio.picascia@unimi.it

Stefano Montanelli
stefano.montanelli@unimi.it

Alfio Ferrara
alfio.ferrara@unimi.it

Nina Tahmasebi
nina.tahmasebi@gu.se

[1] Department of Computer Science, University of Milan, Milan, Italy

[2] Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, Gothenburg, Sweden