Studying Word Meaning Evolution through Incremental Semantic Shift Detection: A Case Study of Italian Parliamentary Speeches

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

[1]University of Milan
[2]Affiliation not available

October 31, 2023

Abstract

The study of semantic shifts, 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 shifts. 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 WiDiD. WiDiD leverages scalable and evolutionary clustering of contextualised word embeddings to detect semantic shifts 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 at least comparable to state-of-the-art approaches, while outperforming the state-of-the-art for certain languages.

Studying Word Meaning Evolution through Incremental Semantic Shift Detection: A Case Study of Italian Parliamentary Speeches

Francesco Periti*, Sergio Picascia**, Alfio Ferrara, Stefano Montanelli, and Nina Tahmasebi

***Abstract*—The study of *semantic shifts*, 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 shifts. 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 WiDiD. WiDiD leverages scalable and evolutionary clustering of contextualised word embeddings to detect semantic shifts 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 at least comparable to state-of-the-art approaches, while outperforming the state-of-the-art for certain languages.

***Index Terms*—Lexical Semantic Change, Semantic Shift Detection, Contextualised Word Embeddings**

I. Introduction

Words are malleable and their meaning(s) continuously evolve, influenced by social practices, events, and political circumstances (Azarbonyad et al., 2017 [1]). 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 [2]). Traditionally, linguists and other scholars in the humanities and social sciences have studied semantic shifts 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].

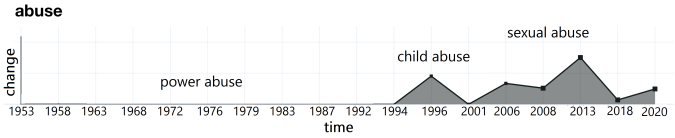

Fig. 1. Change degree of the word "abuso" (i.e., *abuse*) in a diachronic corpus of Italian parliamentary speeches and the evolution of its individual senses. Change is captured using the WiDiD approach presented in Periti et al., 2022 [8]. Before 1994, there is no change and only one sense nodule, *power abuse*; thereafter we observe changes brought about by the emergence of two more sense nodules, namely *child abuse* and *sexual abuse*

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 shifts (Montanelli and Periti, 2023 [5]; Tahmasebi et al., 2021 [6]; Kutuzov et al., 2018 [7]). 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 considered.

Thus far, the research community has concentrated on a simplified SSD task involving mainly change between two time periods (Zamora et al., 2022 [9], Kutuzov et al., 2021 [10]; Basile et al., 2020 [11]; Schlechtweg et al., 2020 [12]). 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

*: Primary contribution, corresponding author
**: Significant contribution
The remaining authors are listed in alphabetical order

The authors are with the Department of Computer Science, University of Milan, Milan, Italy, and also with the Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, Gothenburg, Sweden (e-mail: francesco.periti@unimi.it; sergio.picascia@unimi.it; alfio.ferrara@unimi.it; stefano.montanelli@unimi.it; nina.tahmasebi@gu.se).

Our data and code will be available at https://github.com/FrancescoPeriti/WiDiD

[1]Semantic shift is also often referred to as lexical semantic change, semantic change, as well as sense evolution (Geeraerts, 2020 [3]; Bloomfield, 1994 [4]).

of a dynamic corpus where time documents are progressively added (e.g., *posts* from social networks, Noble et al., 2021 [13]), 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 need 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 (Tahmasebi and Dubossarsky, 23 [14]; Montariol et al., 21 [2]; Kanjirangat et al., 20 [15]). To address this problem, we recently proposed an incremental approach to SSD named WiDiD that enables the analysis of clusters over time (Periti et al., 2022 [16]). In our previous work, we evaluated WiDiD against reference benchmarks for Latin and English on the Graded Change Detection task [12]. This task consists of ranking a set of target words according to their degree of change between two time intervals.

In this paper, WiDiD is extended with a novel cluster analysis to describe the evolution of word meanings over time. In addition, we present a case study where we apply the analysis to a large corpus of Italian parliamentary speeches spanning eighteen different time periods (i.e., eighteen legislatures). Finally, we evaluate WiDiD on seven benchmark datasets.

The remainder of the paper is organised as follows. In Section II, we review the relevant literature on the use of contextualised embeddings for SSD. In Section III, we introduce the WiDiD approach along with the notation that will be used throughout the paper. The novel cluster analysis to describe semantic shift and word meaning evolution is presented in Section IV. A concrete application of these techniques and metrics is illustrated in Section V. The results of WiDiD on the Grade Change Detection task are evaluated in Section VI. Finally, Section VII contains our concluding remarks.

II. Related work

While approaches based on static embeddings are effective in identifying semantic shifts (Tahmasebi et al., 2021 [6]; Kutuzov et al., 2018 [7]), 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 (Montanelli and Periti, 2023 [5]; Periti, 2023 [17]; Periti and Dubossarsky, 2023 [18]; Cassotti et al., 2023 [19]). 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 (Martinc et al., 2020 [20]; Giulianelli et al., 2020 [21]). 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.

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. Shifts 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 time periods are often compared using the Jensen-Shannon divergence criterion (JSD) [21].

Initially, Hu et al., 2019 [22] 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 [21]), HDBSCAN (e.g., Rother et al., 2020 [23]), or Affinity Propagation (e.g., Martinc et al., 2020 [24]) 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 [25]; Tahmasebi and Dubossarsky, 2023 [14]). 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 [26]) - 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 [15] and Montariol et al., 2020 [2] 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 re-aligned. Inspired by Tahmasebi and Risse, 2017 [27], we have recently proposed a novel incremental approach to lexical semantic change, that we have named WiDiD.

TABLE I
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	Polisemy of the word w at time t
$\mathcal{S}^t_w$	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

III. 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 (Periti et al., 2022 [16]). 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 allows assessment of the quantity of semantic shift as well as reconstruction of the evolution of a word's meanings.

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 our previous work, 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 Figure 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.

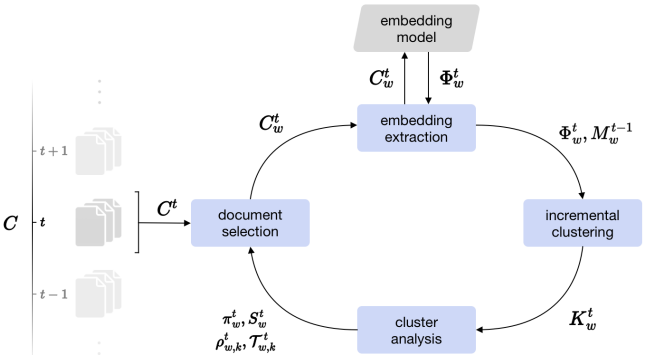

Fig. 2. WiDiD: an incremental approach to Semantic Shift Detection.

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 I.

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 [28]), we consider any derived form of the lemma of w (e.g., plural) as an occurrence of w.

Embedding Extraction (EE): In this step, WiDiD represents each occurrence of the target word w in C^t_w with a different contextualised embedding. The embeddings for w are generated by using a BERT model (Devlin et al., 2019 [29])[2]. The final output of this step is the set Φ^t_w containing all the embeddings of the word w generated for the corpus C^t. Formally,

$$\Phi^t_w = \{e^t_{w,1}, \ldots, e^t_{w,m}\} \ ,$$

where $e^t_{w,j}$ is the contextualised embedding of w in the j-th document and m is the number of documents in C^t_w.

Incremental Clustering (IC): WiDiD first ($t = 0$) uses the standard affinity propagation (AP) algorithm over Φ^0_w (Frey and Dueck, 2007 [30]). This results in a set of clusters denoted as K^0_w.

For $t > 0$, clustering is performed using the A Posteriori affinity Propagation (APP) algorithm proposed in [16] to cluster the embeddings Φ^t_w in groups representing different word meanings (i.e., *sense nodules*). We denote the set of resulting clusters as K^t_w. At each time step, APP creates an additional *sense prototype* embedding $\mu^{t-1}_{w,k}$ for each cluster $k \in K^{t-1}_w$ by averaging all its enclosed embeddings, meaning that $\mu^{t-1}_{w,k}$ 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^{t-1}_w the set of sense prototypes $\mu^{t-1}_{w,k}$ available at time $t-1$. Hence, APP consists

[2]Note that BERT can be replaced with any other contextualised model.

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[3].

Incremental clustering represents a significantly more scalable solution than existing approaches (Montariol et al., 2021 [2]; Kanjirangat et al., 2020 [15]). 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.

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 IV, where we introduce a comprehensive set of metrics specifically designed to describe both a target word and its sense nodules over time.

IV. 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 [25]), i.e., a sense nodule.

A. 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.

[3]A journal paper describing the formalisation and validation of APP in evolutionary clustering scenarios has recently been submitted. Further details are provided in Periti et al., 2022. [16]

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 present at time t. 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.

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

B. 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 [12], $\mathcal{S}_w$ aims to capture the acquisition of a new sense nodule or the loss of an outdated sense nodule.

Following Giulianelli et al., 2020 [21], 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 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 $\mu^{t-1}_{w,k}$.

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

C. 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 shifts presented in Kazi et al., 2022 [31], 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 polisemy π^t_w 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^t_w 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.

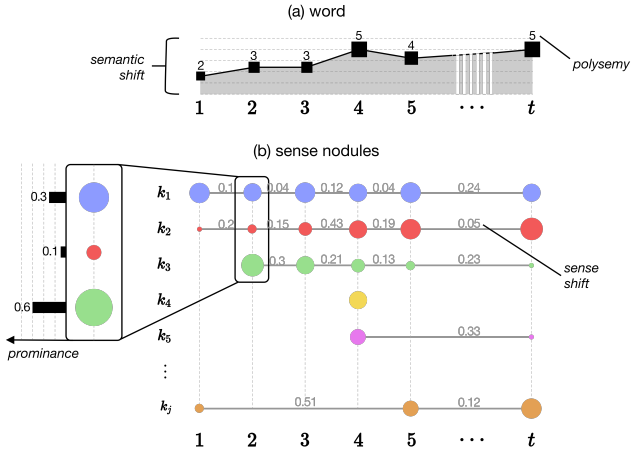

Fig. 3. Clustering visualisation: prototype visualisation of word meaning evolution. Subfigure (a) represents the polisemy and semantic shift of a word over time. Subfigure (b) represents the prominence and sense shift of the sense nodules of that word over time.

V. 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 shifts.

A. 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[4].

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 II.

B. 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 Section III.

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 artefacts introduced during the OCR scanning process. To enhance scalability in this study, as in other studies reported in the literature (e.g., Rodina et al., 2020 [32]), 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*[5]) without performing any fine-tuning (Wolf et al., 2020 [33]). To extract contextualised embeddings for a specific target word w, we fed the model

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

[5]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.

TABLE II
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

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 [34]). 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 scikit-learn implementation for the standard AP algorithm (Pedregosa et al., 2011 [35]). 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.

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 [2], 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.

C. 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 shifts as well as their sense nodules with associated prominence and sense shifts, 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 [36]). Thus, for the sake of readability, we normalised the shift scores of our experiment by the maximum shift value we obtained.

As an example, Figure 4 (a) and 4 (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 4 (a) summarises Figure 4 (b), 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 Figure 4 (b) we adopt the ecological view of word change proposed by Hu et al., 2019 [22]. 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 Figure 4, Table III shows the proportion of documents (i.e., prominence) assigned to each sense nodule.

The cluster analysis in Figure 4 (b) captures examples of semantic shifts 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 [22], 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 III, 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 our previous intuition that computing sense prototypes of large sets of embeddings helps to reduce noise (Periti et al., 2022 [16]). Indeed, we observe a negative correlation between sense shift and the number of documents within a given time interval, meaning that the smaller the

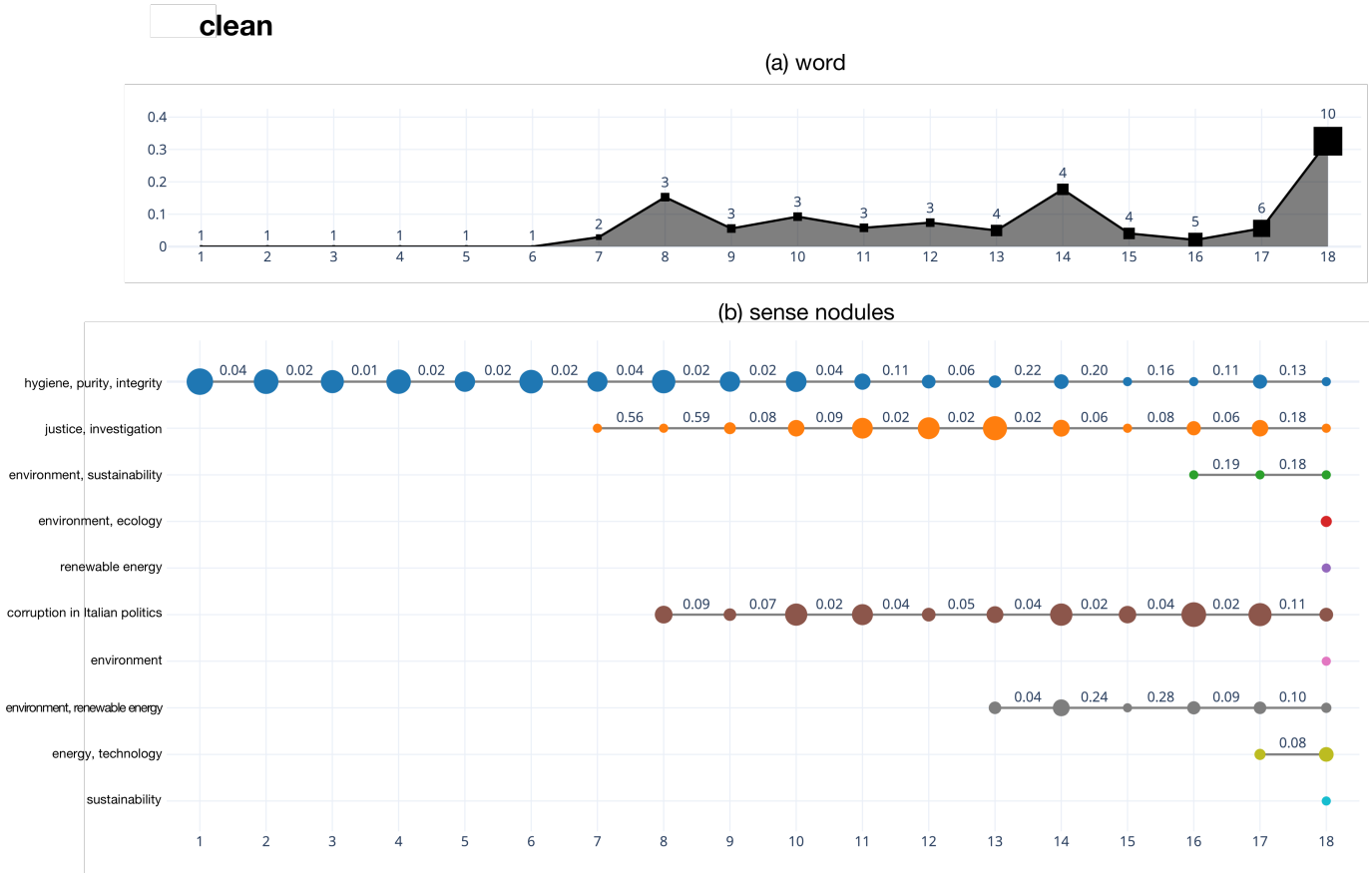

Fig. 4. Clustering visualisation: (a) *semantic shift* and *polisemy* of the Italian word "pulito" (e.g., *clean*); (b) *sense shift* and *prominence* of the sense nodules of the Italian word "pulito" (e.g., *clean*).

TABLE III
PROMINENCE OF THE WORD *clean* OVER TIME. 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

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

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 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 IV and V, respectively.

VI. EVALUATION

In this section, we evaluate the effectiveness and robustness of WiDiD by analysing its performance on various benchmarks

TABLE IV
EXAMPLE OF SEMANTIC SHIFT ASSOCIATED WITH THE CORRESPONDING WORD, TIME INTERVAL, POLISEMY AND A SHORT DESCRIPTION

word	time-interval	polisemy	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 V
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.
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*.

of recent shared tasks such as SemEval-Task 1 (Schlechtweg et al., 2020 [12]), DIACRIta (Basile et al., 2020 [11]), RuShiftEval (Kutuzov et al., 2021 [10]), and LSCDiscovery (Zamora et al., 2022 [9]). 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 [12], 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.

In our previous work [16], we 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 evaluation for RuShiftEval2021 was executed only on Subtask 2, since no binary labels are provided. Furthermore, 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 VI summarises the benchmarks considered.

TABLE VI
PERIOD, SIZE IN TOKENS, REFERENCE, AND NUMBER OF TARGET WORDS FOR THE EVALUATION BENCHMARK CONSIDERED

		Periods	Tokens	Reference	Target Words
SemEval					
English	C_1	1810–1860	6 M	[12]	37
	C_2	1960–2010	6 M		
Latin	C_1	-200–0	65 k	[12]	40
	C_2	0–2000	253 k		
German	C_1	1800–1899	70.2 M	[12]	48
	C_2	1946–1990	72.3 M		
Swedish	C_1	1790–1830	71.0 M	[12]	31
	C_2	1895–1903	110.0 M		
DIACRIta					
Italian	C_1	1945–1970	52 M	[11]	18
	C_2	1990–2014	196 M		
RuShiftEval					
Russian	C_1	1700–1916	94 M	[10]	99
	C_2	1918–1990	123 M		
	C_3	1992–2016	107 M		
LSCDiscovery					
Spanish	C_1	1810–1906	13.0 M	[9]	31
	C_2	1994–2020	22.0 M		

A. Experimental setup

To evaluate WiDiD, we exploited the same setup described in Section V-B 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 Section IV) with the method based on sense nodules recently proposed by Kashleva et al., 2022 [37]). Following [37], 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 [37], we set d as the Canberra distance instead of the cosine distance[6].

In line with previous work (Montanelli and Periti, 2023 [5]), 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[7]. For Subtask 2, we directly used the JSD scores as degree of semantic shift.

B. 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 Table VII and Table VIII, 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 IX 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 VIII 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 [44], 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

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

[7]It is worth noting that, development and training sets are not available for the majority of the benchmark, as LSC is typically framed in an unsupervised scenario (Schlechtweg et al., 2020 [12]). Therefore, the evaluation of Subtask 1 only provides an indication of the model's capability to recognize semantic shifts. 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.

TABLE VII
SUBTASK 2: SPEARMAN'S CORRELATION COEFFICIENTS ACHIEVED FROM VARIOUS STATE-OF-THE-ART EXPERIMENTS. ASTERISKS DENOTE SCORES OBTAINED VIA FINE-TUNING CONTEXTUALISED MODELS, WHILE HYPHENS INDICATE UNAVAILABLE EXPERIMENTAL RESULTS.

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* $C1$-$C3$
Kanjirangatet et al., 2020 [15]	.159	.231	.525	.141	-	-	-	-
Martinc et al., 2020 [38]	.436*	.481	.528*	.238*	-	-	-	-
Karnysheva and Schwarz et al., 2020 [39]	.155	.177	.388	.062	-	-	-	-
Rother et al., 2020 [23]	.306	.321	.605	.268	-	-	-	-
Cuba et al., 2020 [40]	.209	.399	**.656**	.234	-	-	-	-
Montariol et al., 2021 [2]	.456*	**.488***	.561*	**.561***	-	-	-	-
Giulianelli et al., 2022 [41]	.127*	.318*	.287*	-.108*	-	.247*	.267*	.362*
Kashleva et al., 2022 [42]	-	-	-	-	**.553***	-	-	-
WiDiD	**.651**	.433	.527	.499	.544	**.273**	**.393**	**.407**

TABLE VIII
SUBTASK 1: ACCURACY SCORES ACHIEVED FROM VARIOUS STATE-OF-THE-ART EXPERIMENTS. ASTERISKS DENOTE SCORES OBTAINED VIA FINE-TUNING CONTEXTUALISED MODELS, WHILE HYPHENS INDICATE UNAVAILABLE EXPERIMENTAL RESULTS.

References	SemEval				DiacrIta
	English $C1$ - $C2$	*Latin* $C1$ - $C2$	*German* $C1$ - $C2$	*Swedish* $C1$ - $C2$	*Italian* $C1$ - $C2$
Kanjirangatet et al., 2020 [15]	.541	.375	.708	.742	-
Martinc et al., 2020 [38]	.703*	.700	.667*	.710*	-
Karnysheva and Schwarz et al., 2020 [39]	.568	.650	.583	.645	-
Rother et al., 2020 [23]	.622	.575	.729	.742	-
Cuba et al., 2020 [40]	.568	.675	.562	.710	-
Wang et al., 2020 [43]	-	-	-	-	.610*
Giulianelli et al., 2022 [41]	.459*	.500*	.521*	-.516*	.389*
WiDiD	**.757**	**.750**	**.729**	**.774**	**.944**

TABLE IX
EVALUATION SCORES FOR SUBTASK 1 AND SUBTASK 2 ACHIEVED VIA ACCURACY (ACC) AND SPEARMAN'S CORRELATION COEFFICIENTS (CORR), RESPECTIVELY, OVER DIFFERENT BENCHMARKS AND SETUPS. 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.

	JSD / APDP	SemEval				LSCDiscovery	RuShiftEval			DiacrIta
		English $C1$ - $C2$	*Latin* $C1$ - $C2$	*German* $C1$ - $C2$	*Swedish* $C1$ - $C2$	*Spanish* $C1$ - $C2$	*Russian* $C1$ - $C2$	*Russian* $C2$ - $C3$	*Russian* $C1$-$C3$	*Italian* $C1$ - $C2$
Acc Sub. 1	BERT	.622 / .730	-	**.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**	-	.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	-

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).

VII. DISCUSSION AND CONCLUSION

A. 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 et al., 2021 [14]). Furthermore, 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 et al., 2022 [45]; Jiang et al., 2021 [46]). 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 [47]), the authors 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 the functionality of WiDiD but is not meant as an in-depth Social and Linguistics study of the Italian parliament.

B. 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 shifts 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 shifts. 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 shifts 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 while 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. To overcome this issue, an approach that combines WiDiD with global evolutionary clustering can be considered.

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.

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 [22]).

In our specific case study, we used a posteriori analysis. 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.

C. Possible Applications of WiDiD

Both historical linguistics and lexicography involve 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.

D. Concluding remarks

In this paper we have presented WiDiD, the first incremental and scalable approach to Semantic Shift Detection based on the evolutionary clustering of contextualised word embeddings. We demonstrated the practical application of WiDiD on a diachronic corpus of Italian parliamentary speeches spanning eighteen distinct time periods. Finally, we evaluated the performance of WiDiD over seven popular labelled benchmarks. Our empirical results show that, for certain languages, WiDiD outperforms state-of-the-art approaches, while achieving at least comparable results for other languages. At the same time, WiDiD captures significantly more information, and thus allows for more in-depth analysis of the detected change than existing approaches to semantic shift detection.

Acknowledgements

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

References

[1] H. Azarbonyad, M. Dehghani, K. Beelen, A. Arkut, M. Marx, and J. Kamps, "Words Are Malleable: Computing Semantic Shifts in Political and Media Discourse," in *Proc. of CIKM*. Singapore, Singapore: ACM, 2017, pp. 1509–1518.

[2] S. Montariol, M. Martinc, and L. Pivovarova, "Scalable and Interpretable Semantic Change Detection," in *Proc. of NAACL-HLT*. Online: ACL, 2021, pp. 4642–4652.

[3] D. Geeraerts, "Semantic Change: "What The Smurf?"," *The Wiley Blackwell Companion to Semantics*, pp. 1–24, 2020.

[4] L. Bloomfield, *Language*. Motilal Banarsidass Publ., 1994.

[5] S. Montanelli and F. Periti, "A Survey on Contextualised Semantic Shift Detection," 2023.

[6] N. Tahmasebi, L. Borin, and A. Jatowt, "Survey of Computational Approaches to Lexical Semantic Change Detection," 2021.

[7] A. Kutuzov, L. Øvrelid, T. Szymanski, and E. Velldal, "Diachronic Word Embeddings and Semantic Shifts: a Survey," in *Proc. of ICCL*. Santa Fe, New Mexico, USA: ACL, 2018.

[8] S. Castano, A. Ferrara, S. Montanelli, and F. Periti, "Semantic Shift Detection in Vatican Publications: a Case Study from Leo XIII to Francis," in *Proc. of SEBD*. Pisa, Italy: CEUR-WS, 2022, pp. 231–243.

[9] F. D. Zamora-Reina, F. Bravo-Marquez, and D. Schlechtweg, "LSCDiscovery: A Shared Task on Semantic Change Discovery and Detection in Spanish," in *Proc. of LChange*. Dublin, Ireland: ACL, 2022, pp. 149–164.

[10] A. Kutuzov and L. Pivovarova, "RuShiftEval: A Shared Task on Semantic Shift Detection for Russian," in *Proc. of Dialogue*. Online: Redkollegija sbornika, 2021.

[11] P. Basile, A. Caputo, T. Caselli, P. Cassotti, and R. Varvara, "DIACR-Ita@ EVALITA2020: Overview of the EVALITA2020 DiachronicLexical Semantics (DIACR-Ita) Task," in *Proc. of EVALITA*. Online: CEUR-WS, 2020.

[12] D. Schlechtweg, B. McGillivray, S. Hengchen, H. Dubossarsky, and N. Tahmasebi, "SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection," in *Proc. of SemEval*. Barcelona, Spain: ICCL, 2020, pp. 1–23.

[13] B. Noble, A. Sayeed, R. Fernández, and S. Larsson, "Semantic shift in social networks," in *Proc. of SEM*. Online: ACL, 2021, pp. 26–37.

[14] N. Tahmasebi and H. Dubossarsky, "Computational Modeling of Semantic Change," 2023.

[15] V. Kanjirangat, S. Mitrovic, A. Antonucci, and F. Rinaldi, "SST-BERT at SemEval-2020 Task 1: Semantic Shift Tracing by Clustering in BERT-based Embedding Spaces," in *Proc. of SemEval*. Barcelona, Spain: ICCL, 2020, pp. 214–221.

[16] F. Periti, A. Ferrara, S. Montanelli, and M. Ruskov, "What is Done is Done: an Incremental Approach to Semantic Shift Detection," in *Proc. of LChange*. Dublin, Ireland: ACL, 2022, pp. 33–43.

[17] P. Francesco, "Contextualised Semantic Shift Detection," in *Proceedings of the 31st Symposium of Advanced Database Systems (SEBD)*. Galzingano Terme, Italy: CEUR.org, July 2023, pp. 735 – 741. [Online]. Available: https://ceur-ws.org/Vol-3478/paper81.pdf

[18] F. Periti and H. Dubossarsky, "The Time-Embedding Travelers at WiC-ITA," in *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR. org, Parma, Italy*. Parma, Italy: CEUR.org, 2023. [Online]. Available: https://ceur-ws.org/Vol-3473/paper47.pdf

[19] P. Cassotti, L. Siciliani, M. DeGemmis, G. Semeraro, and P. Basile, "XL-LEXEME: WiC Pretrained Model for Cross-Lingual LEXical sEMantic changE," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 1577–1585. [Online]. Available: https://aclanthology.org/2023.acl-short.135

[20] M. Martinc, P. Kralj Novak, and S. Pollak, "Leveraging Contextual Embeddings for Detecting Diachronic Semantic Shift," in *Proc. of LREC*. Marseille, France: ELRA, 2020, pp. 4811–4819.

[21] M. Giulianelli, M. Del Tredici, and R. Fernández, "Analysing Lexical Semantic Change with Contextualised Word Representations," in *Proc. of ACL*. Online: ACL, 2020, pp. 3960–3973.

[22] R. Hu, S. Li, and S. Liang, "Diachronic Sense Modeling with Deep Contextualized Word Embeddings: An Ecological View," in *Proc. of ACL*. Florence, Italy: ACL, 2019, pp. 3899–3908.

[23] D. Rother, T. Haider, and S. Eger, "CMCE at SemEval-2020 Task 1: Clustering on Manifolds of Contextualized Embeddings to Detect Historical Meaning Shifts," in *Proc. of SemEval*. Barcelona, Spain: ICCL, 2020, pp. 187–193.

[24] M. Martinc, S. Montariol, E. Zosa, and L. Pivovarova, *Capturing Evolution in Word Usage: Just Add More Clusters?* New York, NY, USA: ACM, 2020, pp. 343–349.

[25] A. Kutuzov, E. Velldal, and L. Øvrelid, "Contextualized embeddings for semantic change detection: Lessons learned," in *Proc. of NEJLT*, vol. 8, no. 1, 2022.

[26] D. A. Cruse, *Aspects of the Microstructure of Word Meaning*. Oxford University Press, 2000, ch. 2, pp. 30–51.

[27] N. Tahmasebi and T. Risse, "Finding Individual Word Sense Changes and their Delay in Appearance," in *Proc. of RANLP*. Varna, Bulgaria: INCOMA Ltd, 2017, pp. 741–749.

[28] A. Kutuzov, L. Pivovarova, and M. Giulianelli, "Grammatical Profiling for Semantic Change Detection," in *Proc. of CoNLL*, Online, Nov. 2021, pp. 423–434.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. of NAACL-HLT*. Minneapolis, Minnesota: ACL, 2019, pp. 4171–4186.

[30] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[31] R. Kazi, A. Amato, S. Wang, and D. Bucur, "Visualisation Methods for Diachronic Semantic Shift," in *Proc. of the Workshop on Scholarly Document Processing*. Gyeongju, Republic of Korea: ACL, 2022, pp. 89–94.

[32] J. Rodina, Y. Trofimova, A. Kutuzov, and E. Artemova, "ELMo and BERT in Semantic Change Detection for Russian," 2020.

[33] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-Art Natural Language Processing," in *Proc. of EMNLP*. Online: ACL, 2020, pp. 38–45.

[34] R. Sennrich, B. Haddow, and A. Birch, "Neural Machine Translation of Rare Words with Subword Units," in *Proc. of ACL*. Berlin, Germany: ACL, 2016, pp. 1715–1725.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[36] K. Ethayarajh, "How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings," in *Proc. of EMNLP-IJCNLP*. Hong Kong, China: ACL, 2019, pp. 55–65.

[37] K. Kashleva, A. Shein, E. Tukhtina, and S. Vydrina, "HSE at LSCDiscovery in Spanish: Clustering and Profiling for Lexical Semantic Change Discovery," in *Proc. of LChange*. Dublin, Ireland: ACL, 2022, pp. 193–197.

[38] M. Martinc, S. Montariol, E. Zosa, and L. Pivovarova, "Discovery Team at SemEval-2020 Task 1: Context-sensitive Embeddings Not Always Better than Static for Semantic Change Detection," in *Proc. of SemEval*. Barcelona, Spain: ICCL, 2020, pp. 67–73.

[39] A. Karnysheva and P. Schwarz, "TUE at SemEval-2020 Task 1: Detecting Semantic Change by Clustering Contextual Word Embeddings," in *Proc. of SemEval*. Barcelona, Spain: ICCL, 2020, pp. 232–238.

[40] A. Cuba Gyllensten, E. Gogoulou, A. Ekgren, and M. Sahlgren, "SenseCluster at SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection," in *Proc. of SemEval*. Barcelona, Spain: ICCL, 2020, pp. 112–118.

[41] M. Giulianelli, A. Kutuzov, and L. Pivovarova, "Do Not Fire the Linguist: Grammatical Profiles Help Language Models Detect Semantic Change," in *Proc. of LChange*. Dublin, Ireland: ACL, 2022, pp. 54–67.

[42] K. Kashleva, A. Shein, E. Tukhtina, and S. Vydrina, "HSE at LSCDiscovery in Spanish: Clustering and Profiling for Lexical Semantic Change Discovery," in *Proc. of LChange*. Dublin, Ireland: ACL, May 2022, pp. 193–197.

[43] B. Wang, E. Di Buccio, and M. Melucci, "University of Padova @ DIACR-Ita," in *Proc. of EVALITA*. Marrakech, Morocco: CEUR-WS, Dec. 2020.

[44] A. Kutuzov and M. Giulianelli, "UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic Change Detection," in *Proc. of SemEval*. Barcelona, Spain: ICCL, 2020, pp. 126–134.

[45] K. Todorov and G. Colavizza, "An Assessment of the Impact of OCR Noise on Language Models," 2022.

[46] M. Jiang, Y. Hu, G. Worthey, R. C. Dubnicek, T. Underwood, and J. S. Downie, "Impact of OCR quality on BERT embeddings in the domain classification of book excerpts," in *Proc. of CHR*, 2021.

[47] N. Tahmasebi, K. Niklas, G. Zenz, and T. Risse, "On the Applicability of Word Sense Discrimination on 201 Years of Modern English," *International Journal on Digital Libraries*, vol. 13, no. 3-4, pp. 135–153, 2013.