Graph-based Clustering for Detecting Semantic Change Across Time and Languages

Xianghe Ma[1] **Michael Strube**[2] **Wei Zhao**[23]

[1]Faculty of Mathematics and Computer Science, Heidelberg University
[2]Heidelberg Institute for Theoretical Studies
[3]Department of Computing Science, University of Aberdeen
xianghe.ma@stud.uni-heidelberg.de michael.strube@h-its.org
wei.zhao@abdn.ac.uk

Abstract

Despite the predominance of contextualized embeddings in NLP, approaches to detect semantic change relying on these embeddings and clustering methods underperform simpler counterparts based on static word embeddings. This stems from the poor quality of the clustering methods to produce sense clusters—which struggle to capture word senses, especially those with low frequency. This issue hinders the next step in examining how changes in word senses in one language influence another. To address this issue, we propose a graph-based clustering approach to capture nuanced changes in both high- and low-frequency word senses across time and languages, including the acquisition and loss of these senses over time. Our experimental results show that our approach substantially surpasses previous approaches in the SemEval2020 binary classification task across four languages. Moreover, we showcase the ability of our approach as a versatile visualization tool to detect semantic changes in both intra-language and inter-language setups. We make our code and data available[1].

1 Introduction

Since the 19th century, language change has been of ongoing scholarly interest in historical linguistics, stemming from a curiosity to understand intricate genealogy of languages through the comparison of linguistic patterns across earlier and later text corpora (Bopp, 1816; Rask, 1818; Whitney, 1892). Up to now, many more curiosities have emerged, including the establishment of empirical principles of language change (Weinreich et al., 1968; Labov, 1972, 1982, 1994, 2010), the justification of hypothetical pathways of language change (Roberts et al., 2012; Breitbarth, 2014; Lehmann, 2015; Breitbarth, 2019), the investigation of ancestral relationships among hundreds of languages (Boas, 1929; Jäger, 2013; Güldemann, 2018), the discovery of linguistic and extralinguistic factors driving language change (Blaxter, 2015), etc.

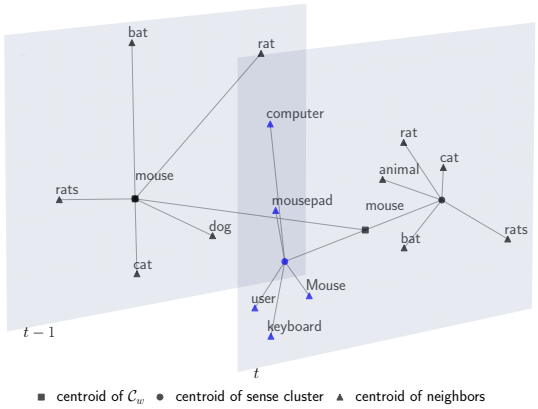


Figure 1: Representation of the semantic changes for 'mouse' in our temporal dynamic graph. Blue nodes indicate the acquisition of a new meaning over time, while black nodes indicate unchanged word meanings.

A seminal work by Coseriu (1970) outlined the characteristics of language change along five dimensions: time, geographical places, medium, registers and social contexts. This has inspired many works to date that leverage these dimensions as a lens to examine changes in grammatical meaning (Traugott, 1985; Bybee and Pagliuca, 1985), syntax (Hale, 1998; Breitbarth, 2022) and many more. In computational linguistics, there has been a surge of interest in leveraging machine learning methods, as cost-efficient alternatives to labor-intensive human inspection. A special focus has been given to detect lexical meaning change, aiming to track changes in word meanings through the analysis of word usages across different time periods (Rohrdantz et al., 2011; Eger and Mehler, 2016; Hamilton et al., 2016a,b,b; Martinc et al., 2020; Gonen et al., 2020; Kaiser et al., 2021; Montariol et al., 2021; Teodorescu et al., 2022; Zamora-Reina et al., 2022).

For instance, Pražák et al. (2020) and Kaiser et al. (2021) leverage static word embeddings to

[1]https://gitlab.com/xiaohaima/lexical-dynamic-graph/

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1542–1561

represent target words across time periods, and then identify the presence of semantic change in each target word by assessing the similarity between its word embeddings from different time periods. Kanjirangat et al. (2020) and Cuba Gyllensten et al. (2020) employ contextualized word embeddings and a clustering method to detect changes in each word sense over time. Although contextualized embeddings excel in many NLP tasks, the performance of these embeddings coupled with clustering methods falls under static counterparts in detecting semantic change (Schlechtweg et al., 2020).

In this work, we identify two major limitations of previous works relying on contextualized embeddings and clustering methods, namely (a) they struggle to capture word senses, especially those with low frequency, leading to poor semantic representation of word senses and (b) they produce time-independent sense clusters and use them to represent word senses varying over time—which is particularly problematic in the presence of a big time gap (e.g., 100 years) between earlier and later time periods. Moreover, these works are limited in scope to detect only intra-language semantic changes. To address these issues, we introduce a graph-based clustering approach that leverages contextualized embeddings to capture the evolution of each word sense across both time and languages. As a result, our approach allows for comparing changes in each word sense across languages over time. This allows for a detailed study of inter-language semantic change, especially to determine if the meanings of word translations across languages remain consistent or diverge over time.

We comparably evaluate our approach in the SemEval2020 binary classification and ranking tasks (Schlechtweg et al., 2020) for detecting semantic change across English, German, Latin and Swedish, and investigate the potential of our graph-based approach, as a visualization tool, to detect semantic changes in both intra-language and inter-language setups. Our findings are summarized below:

- Our approach substantially outperforms the SemEval2020 shared task winner (Pražák et al., 2020) in binary classification across four languages. Our ablation results demonstrate the effectiveness of three crucial components in our approach: our clustering strategy and method, and our distance metric. In the ranking task, our approach performs best in English but falls short in other languages compared to static embedding counterparts.
- We showcase the ability of our approach, as a versatile visualization tool, specifically to (a) track nuanced intra-language semantic changes over time, including both the acquisition and loss of each word sense and (b) track the consistency and divergence of semantic changes over time by comparing detected semantic changes within each language. This aids understanding of inter-language impacts on semantic changes, e.g., new meanings borrowed from other languages.

2 Related Work

Intra-Language Semantic Change Detection. Recently, there has been a growing interest towards detecting meaning changes of target words within each language through a corpus-based study on word usage across time periods (Kutuzov and Giulianelli, 2020; Pömsl and Lyapin, 2020; Giulianelli et al., 2020; Cuba Gyllensten et al., 2020; Karnysheva and Schwarz, 2020; Kaiser et al., 2021; Kutuzov et al., 2022; Card, 2023). Many approaches have been proposed in the SemEval2020 shared tasks (Schlechtweg et al., 2020). Most approaches fall under two categories, based on the choice of word embeddings. For **static embeddings**, approaches, such as Pražák et al. (2020) and Kaiser et al. (2021), begin by refining pre-trained static word embeddings of target words on two corpora from different time periods, resulting in a separate embedding space for each time period. They then employ alignment techniques (Brychcín et al., 2019; Artetxe et al., 2018) to adjust these word embeddings from different time periods. Lastly, a distance measure is applied to these adjusted word embeddings to detect semantic change. For **contextualized word embeddings**, approaches like Kanjirangat et al. (2020) and Cuba Gyllensten et al. (2020) employ the BERT and XLM-R encoders to produce contextualized word embeddings of each target word. They then employ k-means (Rousseeuw, 1987) to partition embeddings of the target word from different time periods into multiple (time-independent) sense clusters. Lastly, a frequency-based criterion is applied to these sense clusters to detect semantic change. Laicher et al. (2021) show that careful data preprocessing can further improve the performance of semantic change detection, and suggest encoding lemmatized target words instead of their original word forms.

Kutuzov et al. (2022) propose to ensemble two top-performing approaches for detecting semantic changes. Kudisov and Arefyev (2022) and Card (2023) propose to detect semantic change by comparing two frequency distributions of a target word across time periods. Each distribution represents the frequencies of vocabulary words predicted as top substitutes for the target word using a masked language model.

Our work differs from others in several aspects: First, we leverage temporal and spatial dynamic graphs, which are derived from BERT embeddings, to represent changes in word meanings across time and space (languages)[2]. This allows for detecting nuanced changes in each word sense, including both the acquisition and loss of meanings over time. Second, we introduce our clustering method and strategy, and our distance metric, designed to produce sense clusters that excel in capturing word senses, especially for low-frequency senses. Moreover, we compute the similarity between sense clusters over time to detect semantic change, while previous approaches do so by applying a frequency-based criterion[3].

Inter-Language Semantic Change Detection. While words in different languages may share a common ancestor and initial meaning, their meanings can diverge over time due to linguistic and extralinguistic variations in these languages. This intriguing phenomenon has led to increased research efforts towards studying semantic changes across languages. To do this, most previous works rely on semantic false friends, namely a pair of words in different languages that share an etymological origin but differ greatly in word meaning (Inkpen et al., 2005; Nakov et al., 2009; Chen and Skiena, 2016; St Arnaud et al., 2017; Uban et al., 2019, 2021). For instance, Uban et al. (2019) employ cross-lingual word embeddings to identify false friends, specifically to determine whether the current meanings of these word pairs have changed. Uban et al. (2021) extended this idea by investigating cross-lingual semantic change laws, specifically by examining the meaning divergence of cognate words from their shared etymological origin.

Furthermore, Montariol and Allauzen (2021) explored bilingual semantic divergence by comparing the meaning changes of mutual word translations of English and French using multilingual BERT. In contrast to our work, they only consider high-frequency word senses, and do not differentiate between the acquisition and loss of meanings over time. Moreover, they do not provide a visual tool to detect semantic divergence across languages.

3 Our Approach

3.1 Semantic-Tree Representation

For each word w, let $\mathcal{C}_w = \{c_1, c_2, \ldots, c_n\}$ be a word cloud consisting of a set of d-dimensional contextualized word embeddings, where n represents the word occurrence in a corpus. We let $e_w \in \mathbb{R}^d$ denote the centroid of $\mathcal{C}_w$, given by $e_w = \frac{1}{n}\sum_i^n c_i$. For any two word clouds, the distance between their centroids e_i and e_j is denoted by $d(e_1, e_2) = 1 - \text{sim}(e_1, e_2)$, where $\text{sim}(e_1, e_2)$ is the cosine similarity between the centroids.

Each word w may exhibit polysemy, manifesting different meanings depending on the context. Therefore, we partition $\mathcal{C}_w$ into m sense clusters, i.e., $\mathcal{C}_w = \bigcup_{1\leq i\leq m} \mathcal{C}_w(p_i)$. Each cluster $\mathcal{C}_w(p_i)$, which is a subset of $\mathcal{C}_w$ centered at p_i, represents a distinct meaning of the polysemous word. For each word, we let $\mathcal{P}_w = \{p_1, \ldots, p_m\}$ denote a set of centroids corresponding to m sense clusters. These centroids are determined using our clustering method (see §3.4). As illustrated in Figure 2, we define a semantic-tree graph that captures multiple recorded meanings of a polysemous word w. We consider the root node e_w, the centroid of $\mathcal{C}_w$, as the **representative embedding**[4] of the word w reoccurring in a corpus. The root node is connected to three nodes on the second layer, which are three sense clusters' centroids $\{p_1, p_2, p_3\}$. We refer to the nodes on the third layer as the representative embeddings of three semantically nearest neighboring words to each centroid p_i.

3.2 Temporal Dynamics within Semantics

We add a temporal dimension to our semantic-tree graph for capturing meaning changes over time. To do this, we denote $\mathcal{C}_w^{t-1}$ and $\mathcal{C}_w^t$ as two point clouds of the word w at two consecutive time periods $t-1$ and t. We then define a temporal dynamic graph

[2]Unlike Schlechtweg et al. (2021), which proposed human-annotated diachronic word usage graphs, our graphs are machine-generated through BERT and additionally offer visual clues regarding meaning changes over time.

[3]Frequency-based criterion: Sense change is detected when a time-independent sense cluster has fewer than 2 tokens in the earlier corpus and more than 5 tokens in the later.

[4]We represent graph nodes as 2-dimensional embeddings, resulting from the PCA projection of high-dimensional contextualized word embeddings.

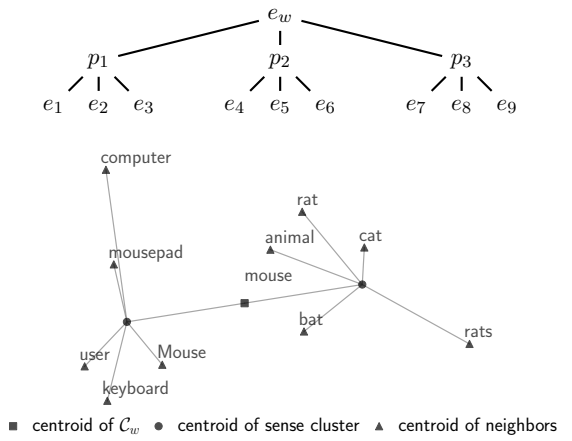


Figure 2: (Top) Representation of polysemous meanings of a word w in a semantic-tree graph. (Bottom) Graph representation of 'mouse' as the root node, generated by applying our approach to the English Wikipedia corpus.

to capture meaning shifts of the word w over time, as illustrated in Figure 3. Given such a graph, we introduce our approach to detect changes in word meaning over time from $t-1$ to t through a two-step process: (a) computing the similarity between sense clusters via our neighbor-based distance metric and (b) utilizing our detection criterion to detect the presence of semantic change.

Bipartite Matching Between Sense Clusters. Our goal is to measure the similarity between a pair of sense clusters from different time periods, i.e., $C_w^{t-1}(p_i^{t-1})$ and $C_w^t(p_j^t)$. This is achieved by measuring the similarity between the centroids of these sense clusters, i.e., p_i^{t-1} and p_j^t, using a bipartite matching method and our **neighbor-based distance metric**[5].

To do this, we define $U_w = \{e_1^{t-1}, \ldots, e_k^{t-1}\}$ as a set of the representative embeddings of k-nearest neighboring words to p_i^{t-1} at time $t-1$. Each e_i^{t-1} is the average of contextual word embeddings of a neighboring word. Similarly, we define $V_w = \{e_1^t, \ldots, e_k^t\}$ as their counterparts to p_j^t at time t. Our bipartite matching problem is given by:

[5]This metric leverages the participation of neighbors to determine the similarity between two words. By doing so, the similarity between two words is less affected by their embedding quality. See Figure 8 (appendix) for the idea illustration.

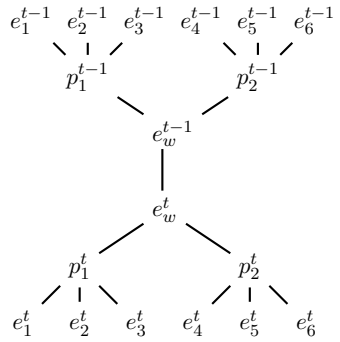


Figure 3: Representation of semantic changes over time in a temporal dynamic graph.

$$\min_{\mu\in\{0,1\}^{k^2}} \sum_{e^{t-1}\in U_w} \sum_{e^t\in V_w} \mu(e^{t-1}, e^t) d(e^{t-1}, e^t)$$
$$\text{s.t.} \sum_{e^t\in V_w} \mu(e^{t-1}, e^t) = 1, \quad \forall e^{t-1} \in U_w$$
$$\sum_{e^{t-1}\in U_w} \mu(e^{t-1}, e^t) = 1, \quad \forall e^t \in V_w$$
$$\mu(e^{t-1}, e^t) \in \{0, 1\}, \quad \forall e^{t-1} \in U_w, e^t \in V_w$$

where $\mu(e^{t-1}, e^t)$ denotes a binary variable that indicates whether a match exists between the input arguments, and $d(e^{t-1}, e^t)$ represents the cosine distance between them. Lastly, the similarity between p_i^{t-1} and p_j^t is given by:

$$s(p_i^{t-1}, p_j^t) = 1 - \frac{1}{|U_w|} \sum_{e^{t-1}\in U_w} \sum_{e^t\in V_w} \hat{\mu}(e^{t-1}, e^t) d(e^{t-1}, e^t)$$

where $\hat{\mu}$ is the optimal solution for bipartite matching, solved by using the Jonker-Volgenant algorithm (Crouse, 2016).

Semantic Change Detection. Our goal is to identify meaning changes over time, especially to distinguish between the acquisition and loss of meanings. We now introduce our **detection criterion**:

For each word w, we denote $M_w = \{p_1^t, \ldots, p_m^t\}$ as a set of the centroids of m sense clusters at time t, and $N_w = \{p_1^{t-1}, \ldots, p_n^{t-1}\}$ as counterparts of n sense clusters at time $t-1$. We then compute the pairwise similarities between the two sets, yielding a semantic similarity matrix denoted below:

$$\mathcal{S} = \begin{pmatrix} s(p_1^{t-1}, p_1^t) & \ldots & s(p_1^{t-1}, p_m^t) \\ \vdots & \ddots & \vdots \\ s(p_n^{t-1}, p_1^t) & \ldots & s(p_n^{t-1}, p_m^t) \end{pmatrix}$$

Based on this matrix, we introduce a threshold t_{sc} to differentiate between acquiring new meanings and losing existing ones:

- If the word at time t loses an existing meaning that it had at time $t-1$, namely p_i^{t-1}, then one cannot find any sense cluster centroids at time t that are similar to p_i^{t-1}. This means that the similarity scores of $s(p_i^{t-1}, p_j^t)$ for all j should fall below t_{sc}, i.e., all the entries in the i-th row of $\mathcal{S}$ are lower than t_{sc}.

- If a word gains a new meaning at time t, i.e., p_j^t, the $s(p_i^{t-1}, p_j^t)$ scores for all i should fall below t_{sc}, i.e., all the entries in the j-th column of $\mathcal{S}$ are lower than t_{sc}.

In either way, be it acquisition or loss of meanings, semantic change is detected.

3.3 Temporal and Spatial Dynamics

Here we extend our temporal dynamic graph to a cross-lingual setup, allowing us to detect semantic change across languages, especially to investigate whether the meanings of word translations across languages change consistently or diverge over time. To do this, we first introduce a spatial dynamic graph, and then combine it with a temporal dynamic graph for detecting/comparing semantic changes over time across languages.

Spatial Dynamic Graph. We let w^{ℓ_1} and w^{ℓ_2} be a pair of mutual word translations in languages ℓ_1 and ℓ_2. Denote $\mathcal{C}_w^{\ell_1}$ as a word cloud consisting of a set of the contextualized word embeddings of the word w^{ℓ_1}, and $e_w^{\ell_1}$ as the word cloud's centroid, $\mathcal{C}_w^{\ell_1}(p_i^{\ell_1})$ as each sense cluster centered at $p_i^{\ell_1}$, and $U_w^{\ell_1} = \{e_1^{\ell_1}, \ldots, e_k^{\ell_1}\}$ as a set of the representative embeddings of k-nearest semantic neighboring words to $p_i^{\ell_1}$. Similarly, we denote $\mathcal{C}_w^{\ell_2}$, $e_w^{\ell_2}$, $p_j^{\ell_2}$, and $V_w^{\ell_2} = \{e_1^{\ell_2}, \ldots, e_k^{\ell_2}\}$ as the counterparts in language ℓ_2. Such a graph is depicted in Figure 4.

We extend the idea of our bipartite matching method to a cross-lingual setup by leveraging k-nearest semantic neighbors $U_w^{\ell_1}$ and $V_w^{\ell_2}$ to compute the similarity between two sense clusters' centroids ($p_i^{\ell_1}$ and $p_j^{\ell_2}$) in different languages. Our bipartite matching problem is adjusted to:

$$\min_{\mu \in \{0,1\}^{k^2}} \sum_{e^{\ell_1} \in U_w^{\ell_1}} \sum_{e^{\ell_2} \in V_w^{\ell_2}} \mu(e^{\ell_1}, e^{\ell_2}) d(e^{\ell_1} + b, e^{\ell_2})$$

where b is a rectified vector that addresses the misalignment between the embedding spaces of languages ℓ_1 and ℓ_2, assuming that one can translate one space to another by using this vector[6]. We refer this vector to the difference between an average token embedding of all words in ℓ_1 and its counterpart in ℓ_2 (Liu et al., 2020).

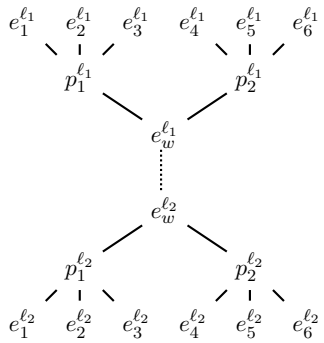


Figure 4: Representation of polysemous meanings of a mutual word translation pair in a spatial dynamic graph.

Once the optimal solution $\hat{u}$ is determined, the similarity between $p_i^{\ell_1}$ and $p_j^{\ell_2}$ is given by:

$$s(p_i^{\ell_1}, p_j^{\ell_2}) = 1 - \frac{1}{|U_w^{\ell_1}|} \sum_{e^{\ell_1} \in U_w} \sum_{e^{\ell_2} \in V_w} \hat{\mu}(e^{\ell_1}, e^{\ell_2}) d(e^{\ell_1} + b, e^{\ell_2})$$

Combining Spatial and Temporal Dynamic Graphs. By adding a temporal dimension to our spatial dynamic graph, the resulting graph captures semantic changes of a mutual word translation pair $e_w^{\ell_1}$ and $e_w^{\ell_2}$ in languages ℓ_1 and ℓ_2 over time from $t-1$ to t—see Figure 9 (appendix).

To detect and compare semantic changes in $e_w^{\ell_1}$ and $e_w^{\ell_2}$, we undertake a two-fold process: For each language, we employ a similarity matrix across sense clusters to detect the acquisition and loss of meanings in $e_w^{\ell_1}$ and $e_w^{\ell_2}$ over time, and then compare the detected changes along two dimensions:

- Consider that $e_w^{\ell_1}$ gains a new meaning $p_i^{\ell_1,t}$ at time t, $e_w^{\ell_2}$ another $p_j^{\ell_2,t}$. If the semantic similarity, given by $s(p_i^{\ell_1,t}, p_j^{\ell_2,t})$, is greater than a cross-lingual threshold t_{cs}, then $e_w^{\ell_1}$ and $e_w^{\ell_2}$ are said to gain a new and similar meaning over time, thereby undergoing consistent acquisition changes in languages ℓ_1 and ℓ_2. Otherwise, their meaning changes over time diverge across languages.

[6] We note that alignments would not affect the results of sense clusters in both source and target languages, as they only shift the embedding space of one language using a translation vector—which does not change the internal structure (topology) of the embedding space.

- Consider that $e_w^{\ell_1}$ at time t loses an existing meaning $p_i^{\ell_1,t-1}$, $e_w^{\ell_2}$ another $p_j^{\ell_2,t-1}$. If $s(p_i^{\ell_1,t-1}, p_j^{\ell_2,t-1}) > t_{cs}$, then $e_w^{\ell_1}$ and $e_w^{\ell_2}$ lose a similar meaning over time, thus undergoing consistent loss changes. Otherwise, their meaning changes differ across languages.

3.4 Our Clustering Method

For each word, our goal is to partition a set of its contextualized word embeddings into multiple sense clusters. To do this, we experimented with the popular k-means method widely adopted in previous works to produce sense clusters; however, we found that this method often produces poor sense clusters that fail to capture word senses, particularly problematic when dealing with low-frequency word senses (see Figure 6). To address this, we present a clustering method that initializes each embedding as a separate cluster. We then iteratively merge two clusters whose centroids are of a distance smaller than a threshold[7] until no further pairs of such similar clusters can be found. The distance between a pair of cluster centroids p_i and p_j is given by $d(p_i, p_j) = \frac{1}{|C_w(p_i)| \cdot |C_w(p_j)|} \sum_{c_i \in C_w(p_i)} \sum_{c_j \in C_w(p_j)} d(c_i, c_j)$. Our method allows for embeddings associated with different word senses (incl. low-frequency senses) to form their own sense clusters. To ensure quality, we exclude clusters with sizes below a threshold, which we consider as noisy clusters. Our method's procedure is provided in Algorithm 1.

In our setup, we iterate through this procedure twice, applying different thresholds t_{sc} and t_{low} each time to achieve specific goals. In the first iteration, we generate a relatively large number of sense clusters for each word. This increases the chance for embeddings with low-frequency word senses to form their own clusters. We then detect and exclude unreliable low-frequency word sense clusters—which we consider as noisy clusters. In the second iteration, we merge non-noisy clusters into only a few to capture word senses of each polysemous word.

4 Experiments

We evaluate our approach in SemEval2020 Task 1 (§4.1), and showcase its ability as a visualization tool to detect semantic changes in both intra-language and inter-language setups (§4.2). We provide analyses regarding our clustering method and embedding spaces—see §A.4 (appendix).

[7] This threshold is tuned on the development sets that we created using ChatGPT. See §A.2 for more details.

Algorithm 1 Our Clustering Method

Require: $\mathcal{C}_w = \{c_i\}_{i=1}^n$ as a set of contextualized embeddings of each word w, t_{sc} as the maximum distance between similar clusters, t_{low} as the minimum cluster size for a low-frequency sense cluster.

Initial centroids of clusters: $\mathcal{P}_w = \{p_i | p_i = c_i\}_{i=1}^n$
while $\min_{p_i \in \mathcal{P}_w, p_j \in \mathcal{P}_w, i \neq j} d(p_i, p_j) < t_{sc}$ **do**
    $\mathcal{P}_w = \mathcal{P}_w \setminus \{p_i, p_j\} \cup \{\frac{p_i + p_j}{2}\}$
end while
for $p_i \in \mathcal{P}_w$ **do**
    if $|\mathcal{C}_w(p_i)| < t_{low}$ **then**
        $\mathcal{P}_w = \mathcal{P}_w \setminus \{p_i\}$
    end if
end for
return $\mathcal{P}_w$

4.1 Intra-language Semantic Change

Setup. We comparably evaluate our approach in SemEval2020 Task 1 for Unsupervised Lexical Semantic Change Detection (Schlechtweg et al., 2020). The task aims to detect intra-language semantic change over time through analyses across two corpora from the 19th and 20th centuries. The task encompasses two subtasks: binary classification and ranking across four languages, i.e., English (EN), German (DE), Latin (LA) and Swedish (SV). We provide data statistics, task descriptions, our implementations details and selection of hyperparameters in §A.2 (appendix). We use the last layer of m-BERT encoder (Devlin et al., 2019) to produce contextualized embeddings of target words across languages on the lemmatized corpora.

Results. Table 1 compares our approach with its counterparts that rely on static and contextualized word embeddings in the SemEval2020 binary classification task (See §A.1 for the results in the ranking task). We find that UWB and LiverLanguage based on static word embeddings outperform NLP@IDSIA and Skurt relying on contextualized embeddings. This unexpected result has been observed previously in Schlechtweg et al. (2020), where the work attributes the underperformance of NLP@IDSIA and Skurt to the fact that they do not sufficiently leverage the power of contextualized embeddings. However, our approach based on contextualized embeddings largely outperforms all others, demonstrating its superiority in leveraging contextualized embeddings. The sources of our improvement are manifold: First, our ap-

Approaches	Avg	EN	DE	LA	SV
Static Word Embeddings					
UWB (Pražák et al., 2020)	.687	.622	.750	**.700**	.677
Life-Language (Asgari et al., 2020)	.686	.703	.750	.550	.742
Contextualized Word Embeddings					
NLP@IDSIA (Kanjirangat et al., 2020)	.637	.622	.625	.625	.677
Skurt (Cuba Gyllensten et al., 2020)	.629	.568	.562	.675	.710
Our Approach	**.776**	**.784**	**.813**	**.700**	**.806**

Table 1: Accuracies from our approach and its counterparts in the SemEval2020 binary classification task.

proach includes a parameterized threshold that we use to stop our clustering process. This threshold is adjusted on the development sets that we created using ChatGPT, while previous approaches lack access to the development sets. Undoubtedly, our approach gains advantages from that, but more importantly, we argue that our improvement results from the careful design of our components—which we demonstrate through an ablation study.

Components	Approaches	EN
All-in-one	Our Approach	.784
Clustering Strategy	$\ominus$ Time-dep. $\oplus$ Time-indep.	.649
Clustering Method	$\ominus$ Our method $\oplus$ k-means	.649
Distance Metric	$\ominus$ Neighbor-based $\oplus$ Euclidean	.676

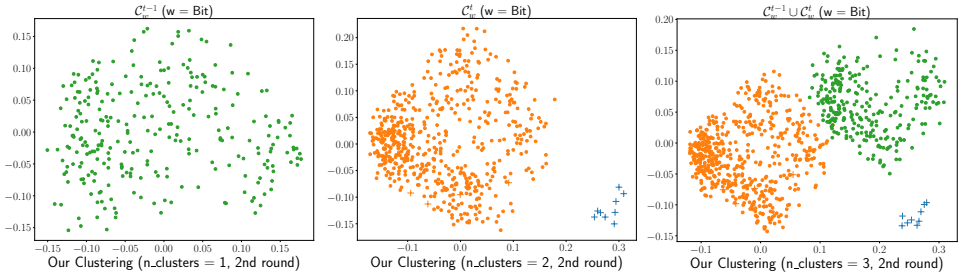
Table 2: Ablation test in the SemEval2020 binary classification task, where $\ominus$ X $\oplus$ Y means the replacement of component X in our approach by component Y.

Ablation Study. Table 2 reports the ablation results on the three crucial components of our approach[8]. **First**, we find that generating time-dependent sense clusters yields much better results than time-independent counterparts adopted in previous works (Kanjirangat et al., 2020; Cuba Gyllensten et al., 2020). We believe the previous approaches are based on the assumption that if a target word remains a consistent meaning over time, its word embeddings from different time periods should be grouped into a single time-independent sense cluster. However, this is challenging due to big context variations between the 19th and 20th corpora, causing contextualized encoders like BERT to misinterpret the consistent meaning as multiple dissimilar senses. As such, time-independent clusters misinterpret these senses over time as spurious meaning changes. To support this hypothesis, we compare our clustering method between the time-independent and time-dependent setups. We see that Figure 5 (a)+(b), which produce sense clusters at each time period, adeptly capture the word senses of 'bit' at each time. However, in Figure 5 (c) there are three distinct sense clusters, and the one in green, which has the same meaning of the orange one, is misinterpreted as a spurious new meaning by contextualized encoders.

Second, we see that our clustering method considerably outperforms the popular k-means. The reasons for this are depicted in Figure 6: (a) with $k = 2$ (too small), most embeddings representing the low-frequency word sense (marked with '+') are wrongly subsumed into the high-frequency sense cluster in blue, and moreover, the two sense clusters in orange and blue share the same meaning and should not be separated; (b) with $k = 8$ (too large): the high-frequency sense is wrongly divided into multiple sense clusters, despite low-frequency sense being correctly identified and mostly forming a distinct cluster in gray; (c) our clustering method produce two sense clusters that effectively capture both high-frequency and low-frequency senses. This is because our approach does not fix the number of clusters but instead leverage the idea of iteratively merging clusters until convergence, subject to some conditions. This provides the flexibility to find an adaptable number of clusters.

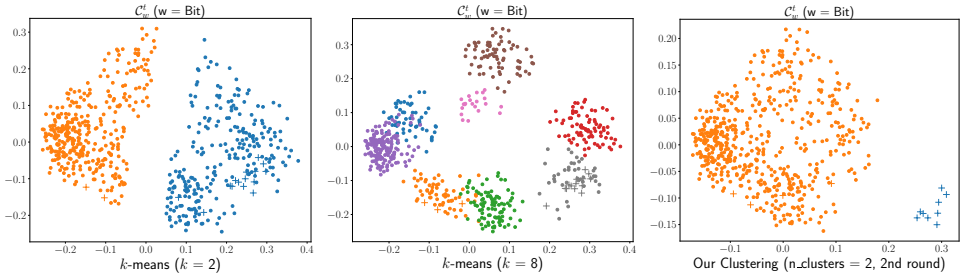
Lastly, we see that neighbor-based distance metric greatly surpasses Euclidean distance. Unlike Euclidean distance, which quantifies the similarity between sense clusters by computing the similarity between cluster centroids, our neighbor-based metric does this by computing the similarity between the k semantically nearest neighbors to each cluster centroid. We believe that our metric, which leverages k neighbors rather than just the centroid, allows us to better capture the semantics of sense clusters, providing a more accurate reflection of the similarity between sense clusters.

[8]We contrast the components of our approach with baseline approaches in Table 10 (appendix).



(a) Time-dependent at time $t-1$ (b) Time-dependent at time t (c) Time-independent

Figure 5: Comparison of sense clusters for the word 'bit' between the time-dependent (at time $t-1$ and t) and time-independent setups. Color indicates the cluster assignment of each point. A dot point represents the high-frequency word sense (a small piece), while a '+' indicates the low-frequency sense (binary digit).



(a) Time-dependent at time t (b) Time-dependent at time t (c) Time-dependent at time t

Figure 6: Comparison of sense clusters produced by k-means and our method.

4.2 Exploratory Study

Our setup is detailed in §A.3 (appendix).

Intra-Language Semantic Change. Consider the polysemous English word 'mouse'. It is well-known that the word's meaning has evolved from a small rodent to a computer input device over time. Figure 1 showcases the ability of our temporal dynamic graph, which is derived from our approach on both historical and present English corpora, to capture the recorded semantic changes of the word 'mouse' over time. We find that the word initially has only one meaning represented by its five-nearest neighbors such as 'rat' and 'bat' at time $t-1$. As time progresses, the word maintains this original meaning while gaining a new meaning about computer device at time t—characterized by its corresponding neighbors in blue color. We find many such examples across languages, and provide a few in Figure 11 (appendix).

Inter-Language Semantic Changes. Figure 7 compares the detected semantic changes for the word translations 'mouse', 'Maus', and 'mus' across English, German and Swedish over time. We observe that each of these word translations holds just a single meaning at time $t-1$, within the 19-century historical corpus. However, at time t within the current Wikipedia corpus, 'mouse' and 'Maus' gain new meanings, indicated in blue nodes, while 'mus' gains two new meanings, similarly indicated. Furthermore, we note that the blue nodes with the same meaning of "computer device" are labeled in orange text across languages: Both 'mouse' and 'Maus' acquire the same new meaning, implying that these two words undergo consistent acquisition changes over time. However, the meaning changes of these two words diverge from that of 'mus': Although all three words acquire the same meaning "computer device", 'mus' gains another new meaning related to 'svans' and 'päls' at time t. This example showcases the potential of our approach as a visualization tool to detect semantic divergence and consistency across languages over time. We provide examples regarding meaning loss in Figure 12 (appendix).

We validated these results on both Wiktionary and the etymological dictionary[9]. While the overall results are accurate, the neighbors connected to each root node do not necessarily represent synonyms of that node. For instance, in Figure 7 (left), both 'cat' and 'dog' are closer than 'rat' to 'mouse'.

[9] https://www.etymonline.com/

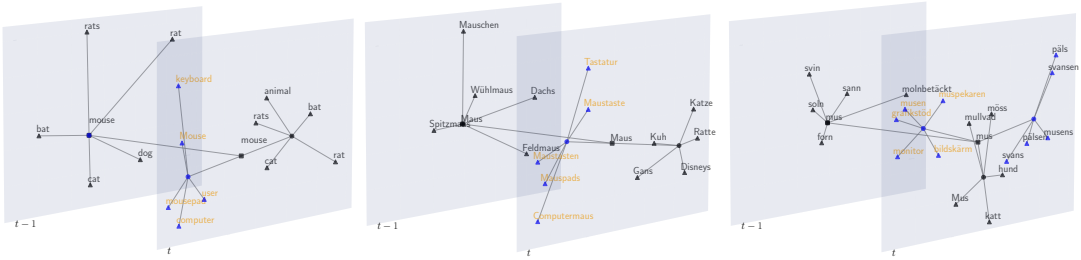


Figure 7: Representation of the detected semantic changes for the word translations 'mouse', 'Maus', and 'mus' in three temporal dynamic graphs. Blue nodes indicate the acquisition of new meanings, while orange text additionally marks certain new meanings that are considered highly similar, i.e., undergoing consistent acquisition change.

This only means that the contexts in which 'cat' and 'dog' appear are more similar to the context of 'mouse'.

5 Conclusions

We proposed a graph-based clustering approach to capture changes within each word sense across both time and languages. We addressed an intriguing concern that contextualized embeddings coupled with clustering methods seem not suitable for detecting semantic change, as they underperform their static embedding counterparts. We identified a crucial reason for this: Previous approaches rely on low-quality clustering methods to handle contextualized embeddings. Our results demonstrated that, when equipped with an appropriate clustering method, strategy, and distance metric, contextualized embeddings can produce high-quality sense clusters that effectively capture word senses, even low-frequency ones. These factors attribute to our approach's superiority over the shared task winner, ULB, which relies on static embedding. Further, the use of our approach as a visualization tool highlights its value in conducting exploratory studies on both intra- and inter-language semantic changes.

6 Limitations

Our approach still lags behind static embedding counterparts in the ranking task for languages other than English (see §A.1). Further improvements may result from improving the quality of embeddings for non-English languages.

Lack of a standard evaluation setup poses a challenge in tracking recent progress in the **intra-language** setup. For instance, Card (2023) reported results in SemEval2020 and GEM ranking tasks; Teodorescu et al. (2022) did in LSCDiscovery binary and ranking tasks; we did in SemEval2020 classification and ranking tasks.

Further, the absence of benchmark datasets for detecting the divergence of semantic changes in the **inter-language** setup poses another challenge in evaluating our approach. Moreover, in the cross-lingual setup, we addressed the misalignment between the embedding spaces of two languages; however, our focus was on word-level rather than meaning-level alignments. Thus, it remains unclear how the embedding spaces (adjusted via word-level alignments) handle words with polysemy profiles. These present avenues for future work.

7 Ethical Considerations

Our work proposed an approach based on BERT to detect semantic change and evaluated the approach on the historical datasets from SemEval2020 Task 1. We acknowledge the potential biases arising from both our approach and the datasets. In historical corpora, a bias towards male authors is often observed. Regarding our approach, BERT is known to encode social biases related to gender and race. Up to now, it remains unclear how these biases may affect the results of semantic change detection. We leave this question to future work.

Acknowledgements

We thank the anonymous reviewers for their thoughtful comments that greatly improved the texts. We also thank Dominik Schlechtweg and Steffen Eger for providing feedback on the early version of this work. This work has been supported by the Klaus Tschira Foundation and Young Marsilius Fellowship, Heidelberg, Germany.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Pro-*

ceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Ehsaneddin Asgari, Christoph Ringlstetter, and Hinrich Schütze. 2020. EmbLexChange at SemEval-2020 task 1: Unsupervised embedding-based detection of lexical semantic changes. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 201–207, Barcelona (online). International Committee for Computational Linguistics.

Tam Blaxter. 2015. Gender and language change in old norse sentential negatives. *Language Variation and Change*, 27(3):349–375.

Franz Boas. 1929. Classification of american indian languages. *Language*, pages 1–7.

Franz Bopp. 1816. *Über das Conjugationssystem der Sanskritsprache in Vergleichung mit jenem der griechischen, lateinischen, persischen und germanischen Sprache: Nebst Episoden des Ramajan und Mahabharat und einigen Abschnitten aus den Vedas*. Andreä.

Anne Breitbarth. 2014. The development of conditional should in english. *Language change at the syntax-semantics interface*, 278:293.

Anne Breitbarth. 2019. Should a conditional marker arise... the diachronic development of conditional sollte in german. *Glossa: a journal of general linguistics*, 4(1).

Anne Breitbarth. 2022. V3 after central adverbials in german: continuity or change? *Journal of historical syntax*.

Tomáš Brychcín, Stephen Taylor, and Lukáš Svoboda. 2019. Cross-lingual word analogies using linear transformations between semantic spaces. *Expert Systems with Applications*, 135:287–295.

Joan L Bybee and William Pagliuca. 1985. Cross-linguistic comparison and the development of grammatical meaning. *Historical semantics, historical word formation*, 59.

Dallas Card. 2023. Substitution-based semantic change detection using contextual embeddings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 590–602, Toronto, Canada. Association for Computational Linguistics.

Yanqing Chen and Steven Skiena. 2016. False-friend detection and entity matching via unsupervised transliteration. *arXiv preprint arXiv:1611.06722*.

Eugenio Coseriu. 1970. *Einführung in die strukturelle Betrachtung des Wortschatzes*, volume 14. Tübinger Beiträge zur Linguistik.

David F Crouse. 2016. On implementing 2d rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696.

Amaru Cuba Gyllensten, Evangelia Gogoulou, Ariel Ekgren, and Magnus Sahlgren. 2020. SenseCluster at SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 112–118, Barcelona (online). International Committee for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Steffen Eger and Alexander Mehler. 2016. On the linearity of semantic change: Investigating meaning variation via dynamic graph models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 52–58, Berlin, Germany. Association for Computational Linguistics.

Brendan J Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *science*, 315(5814):972–976.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. 2020. Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555, Online. Association for Computational Linguistics.

Tom Güldemann. 2018. Historical linguistics and genealogical language classification in africa. *The languages and linguistics of Africa*, (11).

Mark Hale. 1998. Diachronic syntax. *Syntax*, 1(1):1–18.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas. Association for Computational Linguistics.

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

Diana Inkpen, Oana Frunza, and Grzegorz Kondrak. 2005. Automatic identification of cognates and false friends in french and english. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, volume 9, pages 251–257.

Gerhard Jäger. 2013. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Language Dynamics and Change*, 3(2):245–291.

Jens Kaiser, Sinan Kurtyigit, Serge Kotchourko, and Dominik Schlechtweg. 2021. Effects of pre- and post-processing on type-based embeddings in lexical semantic change detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 125–137, Online. Association for Computational Linguistics.

Vani Kanjirangat, Sandra Mitrovic, Alessandro Antonucci, and Fabio Rinaldi. 2020. SST-BERT at SemEval-2020 task 1: Semantic shift tracing by clustering in BERT-based embedding spaces. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 214–221, Barcelona (online). International Committee for Computational Linguistics.

Anna Karnysheva and Pia Schwarz. 2020. TUE at SemEval-2020 task 1: Detecting semantic change by clustering contextual word embeddings. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 232–238, Barcelona (online). International Committee for Computational Linguistics.

Artem Kudisov and Nikolay Arefyev. 2022. BOS at LSCDiscovery: Lexical substitution for interpretable lexical semantic change detection. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 165–172, Dublin, Ireland. Association for Computational Linguistics.

Andrey Kutuzov and Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 126–134, Barcelona (online). International Committee for Computational Linguistics.

Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2022. Contextualized embeddings for semantic change detection: Lessons learned. In *Northern European Journal of Language Technology, Volume 8*, Copenhagen, Denmark. Northern European Association of Language Technology.

W Labov. 1994. Principles of linguistic change, volume 1: Internal factors (p. 664). *Malden, MA: Wiley-Blackwell*.

William Labov. 1972. Some principles of linguistic methodology. *Language in society*, 1(1):97–120.

William Labov. 1982. Building on empirical foundations. *Perspectives on historical linguistics*, 24:17.

William Labov. 2010. Principles of linguistic change. iii: Cognitive and cultural factors. *Malden MA: Wiley-Blackwell*.

Severin Laicher, Sinan Kurtyigit, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Explaining and improving BERT performance on lexical semantic change detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 192–202, Online. Association for Computational Linguistics.

Christian Lehmann. 2015. *Thoughts on grammaticalization*. Language Science Press.

Chi-Liang Liu, Tsung-Yuan Hsu, Yung-Sung Chuang, and Hung-Yi Lee. 2020. A study of cross-lingual ability and language-specific information in multilingual bert. *arXiv preprint arXiv:2004.09205*.

Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020. Capturing evolution in word usage: just add more clusters? In *Companion Proceedings of the Web Conference 2020*, pages 343–349.

Syrielle Montariol and Alexandre Allauzen. 2021. Measure and evaluation of semantic divergence across two languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1247–1258, Online. Association for Computational Linguistics.

Syrielle Montariol, Matej Martinc, and Lidia Pivovarova. 2021. Scalable and interpretable semantic change detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4642–4652, Online. Association for Computational Linguistics.

Svetlin Nakov, Preslav Nakov, and Elena Paskaleva. 2009. Unsupervised extraction of false Friends from parallel bi-texts using the web as a corpus. In *Proceedings of the International Conference RANLP-2009*, pages 292–298, Borovets, Bulgaria. Association for Computational Linguistics.

Martin Pömsl and Roman Lyapin. 2020. CIRCE at SemEval-2020 task 1: Ensembling context-free and context-dependent word representations. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 180–186, Barcelona (online). International Committee for Computational Linguistics.

Ondřej Pražák, Pavel Přibáň, Stephen Taylor, and Jakub Sido. 2020. UWB at SemEval-2020 task 1: Lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 246–254, Barcelona (online). International Committee for Computational Linguistics.

Rasmus Rask. 1818. *Undersøgelse om det gamle nordiske eller islandske Sprogs Oprindelse*. Gyldendal.

Ian Roberts, Laura Brugé, Anna Cardinaletti, Giuliana Giusti, Nicola Munaro, and Cecilia Poletto. 2012. Diachrony and cartography: Paths of grammaticalization and the clausal hierarchy. *Functional heads: The cartography of syntactic structures*, 7:351–367.

Christian Rohrdantz, Annette Hautli, Thomas Mayer, Miriam Butt, Daniel A. Keim, and Frans Plank. 2011. Towards tracking semantic change by visual analytics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 305–310, Portland, Oregon, USA. Association for Computational Linguistics.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021. DWUG: A large resource of diachronic word usage graphs in four languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7079–7091, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Adam St Arnaud et al. 2017. Identifying cognate sets across dictionaries of related languages.

Daniela Teodorescu, Spencer von der Ohe, and Grzegorz Kondrak. 2022. UAlberta at LSCDiscovery: Lexical semantic change detection via word sense disambiguation. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 180–186, Dublin, Ireland. Association for Computational Linguistics.

Elizabeth Traugott. 1985. On regularity in semantic change.

Ana-Sabina Uban, Alina Maria Ciobanu, and Liviu P Dinu. 2021. Cross-lingual laws of semantic change. *Computational approaches to semantic change*, 6:219.

Ana-Sabina Uban, Alina Cristea, and Liviu P Dinu. 2019. A computational approach to measuring the semantic divergence of cognates. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 96–108. Springer.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2009. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th annual international conference on machine learning*, pages 1073–1080.

Uriel Weinreich, William Labov, and Marvin Herzog. 1968. *Empirical foundations for a theory of language change*. University of Texas Press.

William Dwight Whitney. 1892. *Max Müller and the science of language: A criticism*. D. Appleton.

Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. LSCDiscovery: A shared task on semantic change discovery and detection in Spanish. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 149–164, Dublin, Ireland. Association for Computational Linguistics.

Jinan Zhou and Jiaxin Li. 2020. TemporalTeller at SemEval-2020 task 1: Unsupervised lexical semantic change detection with temporal referencing. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 222–231, Barcelona (online). International Committee for Computational Linguistics.

A Appendix

A.1 Ranking Task.

Our approach. We describe our approach used to perform the SemEval2020 ranking task. Following many works (Schlechtweg et al., 2020; Kanjirangat et al., 2020; Kutuzov and Giulianelli, 2020), we design a criterion to grade the degree of semantic change by comparing the frequencies of word meanings across time periods. Such a criterion allows for capturing both past and prospective changes in word meanings. For instance, when comparing the frequencies of a word meaning over time, a frequency decline from time $t-1$ to t suggests the potential loss of the meaning in the future. Here, we aim to measure the degree of both meaning acquisition and loss over time. Our criterion is detailed below:

Once the sets of cluster centroids $\mathcal{P}_w^{t-1}$ and $\mathcal{P}_w^t$ at time $t-1$ and t are determined[10], we then divide the combined set of these centroids into m clusters. Each cluster H_i represents a distinct word sense over time, and comprises centroids from different time periods that exhibit high similarities exceeding a threshold t_{sc}. This implies that the associated meanings of these centroids remain unchanged over time. For each target word, we let $A^t = (a_1^t, \ldots, a_m^t)$ denote the frequency distribution of m word senses at time t, where $a_i^t = \frac{\sum_{p \in H_i \cap \mathcal{P}_w^t} |C_w^t(p)|}{\sum_{p \in \mathcal{P}_w^t} |C_w^t(p)|}$, and similarity for A^{t-1}. By comparing the two frequency distributions, we illustrate three scenarios:

- $a_i^{t-1} = 0$ and $a_i^t > 0$: acquiring a new meaning at time t.
- $a_i^t = 0$ and $a_i^{t-1} > 0$: losing an existing meaning at $t-1$.
- $a_i^t > 0$ and $a_i^{t-1} > 0$: indicating the degree of meaning change over time.

Follow Schlechtweg et al. (2020), we grade the degree of semantic change by computing the Jensen-Shannon distance between two frequency distributions, noted as $\mathrm{JSD}(A^t, A^{t-1})$ in our setup.

Results. Table 3 compares the results of our approach and its counterparts in the SemEval2020 ranking task. We see that our approach performs best among the approaches relying on contextualized word embeddings. Our approach substantially outperforms the recent substitution-based approach in 3 out of 4 languages, and surpasses static embedding counterparts in English. However, our approach still lags behind in other languages; interestingly, it outperforms static embedding counterparts in all languages for binary classification. Our analysis on this is the following:

First, the ranking task is inherently more challenging, as it requires to quantify the fine-grained degree of semantic change. Second, m-BERT is known to produce different embedding quality across languages, with superior embedding quality in English. In binary classification, where the task is straightforward, embedding quality matters little. However, for the challenging ranking task, lower-quality embeddings can harm the results. We leave the verification of this hypothesis to future work.

[10] In contrast to the binary classification setup, our clustering method does not exclude outliers in the ranking setup.

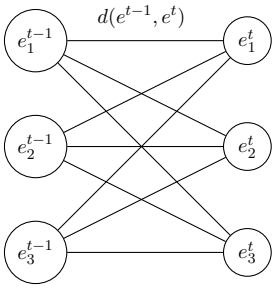


Figure 8: Illustration of bipartite matching for computing the similarity between sense clusters' centroids p_1^t and p_1^{t-1}. Here, $\{e_i^t\}_{i=1}^3$ indicates the representative embeddings of three semantically nearest neighboring words to p_1^t. The same applies to $\{e_i^{t-1}\}_{i=1}^3$ and p_1^{t-1}.

A.2 Experimental Setups for SemEval2020 Task 1

Datasets. Table 4 provides data statistics for the SemEval2020 Task 1.

Task Descriptions. SemEval2020 Task 1 consists of two subtasks, namely (a) binary classification, where one decides whether the meaning of each target word has changed over time by analyzing word usage across two text corpora from different time periods and (b) ranking, where a list of provided target words should be ranked based on scores given by a criterion indicating the degree to which each word undergoes semantic change.

Implementation Details. For each language, we produce contextualized word embeddings of target words from two time periods of text corpora, and

Approaches	Avg	EN	DE	LA	SV
Static Word Embeddings					
UG_Student_Intern (Pömsl and Lyapin, 2020)	**.527**	.422	**.725**	.412	.547
Jiaxin & Jinan (Zhou and Li, 2020)	.518	.325	.717	.440	**.588**
Contextualized Word Embeddings					
Substitution (Card, 2023)	.488	.547	.563	**.533**	.310
Skurt (Cuba Gyllensten et al., 2020)	.374	.209	.656	.399	.234
Our Approach	.506	**.569**	.656	.377	.423

Table 3: Results from our approach and its counterparts in the SemEval2020 ranking task. Results are reported in Spearman correlation.

	Corpus #1					Corpus #2					
Language	period ($t-1$)	#tokens	avg/t	max/t	min/t	period (t)	#tokens	avg/t	max/t	min/t	#targets
English	1810-1860	25,955	701	4,211	86	1960-2010	30,060	812	4,062	106	37
German	1800-1900	71,556	1,490	28,756	35	1946-1990	42,260	880	8,539	103	48
Latin	200BC-1BC	27,548	688	3,498	26	100AD-present	129,568	3,239	10,362	245	40
Swedish	1790-1830	35,021	1,129	6,934	83	1895-1903	126,126	4,068	14,583	89	31

Table 4: Statistics of the SemEval2020 Task 1 Corpus. The last column '#targets' denotes the number of target words, while the column '#tokens' denotes the total token count of target words. The column 'avg/t' indicates the average token count of each target word, while the column 'max/t' indicates the maximum token count per target word, and the column 'min/t' indicates the minimum token count per target word.

then employ our clustering method to partition embeddings of each target word into multiple sense clusters in order to constitute a temporal dynamic graph. As mentioned previously, we iterate through our clustering procedure twice. This requires two chosen hyperparameters in each iteration: (a) t^0_{low} and t^1_{low}, representing the minimum occurrence for a low-frequency meaning, i.e., the minimum cluster size and (b) t^0_{sc} and t^1_{sc}, representing the maximum distance between similar clusters. Furthermore, we need to leverage bipartite matching based on k-nearest semantic neighbors (with k as an extra hyperparameter) to compute the similarity between sense clusters. This step is crucial for detecting nuanced meaning changes over time.

We use a grid search to tune the following two hyperparameters on **the development set we constructed using ChatGPT** in each language: $t^0_{sc} \in \{t|0.10 < t < 0.35, 100 \cdot t \in N_+\}$ and $t^1_{sc} \in \{t|0.10 < t < 0.45, 100 \cdot t \in N_+\}$. In all setups, we set t^0_{low} to 5 and consider clusters with sizes below 5 as noisy clusters[11]. We set t^1_{low} to 0, as noisy clusters should have been removed when the first iteration ends. We set k to 14—see our clustering analysis in §A.4. Our configuration of these hyperparameters across languages are reported in Table 5 and 6 for classification and ranking tasks. We note that the chosen t^1_{sc} is applied to our detection criterion for finding similar and dissimilar sense clusters, i.e., to detect the presence of semantic change in each word sense.

Languages	t^0_{sc}	t^1_{sc}	k	t^0_{low}	t^1_{low}
English	0.34	0.40	14	5	0
German	0.22	0.38	14	5	0
Latin	0.16	0.16	14	5	0
Swedish	0.28	0.32	14	5	0

Table 5: Configuration of hyperparameters across languages in the SemEval2020 binary classification task.

Languages	t^0_{sc}	t^1_{sc}	k	t^0_{low}	t^1_{low}
English	0.34	0.40	14	0	0
German	0.22	0.38	14	0	0
Latin	0.16	0.16	14	0	0
Swedish	0.28	0.32	14	0	0

Table 6: Configuration of hyperparameters across languages in the SemEval2020 ranking task.

Construction of Development Sets using ChatGPT. As SemEval2020 Task 1 operates in an

[11] In the SemEval 2020 shared task, a new word sense is acknowledged upon meeting two rules: (a) this sense associates with fewer than 2 word tokens at time t-1 and (b) it associates with more than 5 word tokens at time t. If a word sense meets (a) but violates (b), for example, having less than 5 word tokens at time t, then this new sense is considered unacceptable and categorized as a noisy sense. We follow this idea and remove sense clusters with word tokens fewer than 5.

	Corpus #1					Corpus #2					
Language	period ($t-1$)	#tokens	avg/t	max/t	min/t	period (t)	#tokens	avg/t	max/t	min/t	#targets
English	1810-1860	3,294	329	947	23	Wiki (08.2023)	53,019	5,301	23,733	755	10
German	1800-1900	17,893	1,789	4,536	84	Wiki (08.2023)	43,963	4,396	22,809	240	10
Swedish	1790-1830	12,409	1,240	5,310	14	Wiki (08.2023)	47,629	4,762	35,249	25	10

Table 7: Statistics of the corpus in the inter-language setup.

unsupervised setting, the task lacks development sets, with the entire corpus treated as evaluation sets. Here, we create a development set per language on which we tune the hyperparameters of our clustering approach for each language. Each development set includes 8 target words that are unseen in evaluation sets. Each target word associates with two senses. We use ChatGPT-3.5 to produce 100 sentences that contextualize each sense. We now describe our data construction approach in detail:

For each target word, we begin by instructing ChatGPT to provide a list of possible word senses, and then verify their accuracy using Wiktionary. After that, we select two verified word senses from the list and instruct ChatGPT to generate a corpus of n sentence that evenly incorporate both word senses of the target word. Then, we construct a gold label vector, denoted as $\mathcal{Y}_w = [y_1, \ldots, y_n]$ with $y_i \in \{0, 1\}$, where y_i specifies whether the target word in the i-th sentence within the corpus corresponds to the first or second word sense.

We observed that ChatGPT yields sentences of satisfactory quality, which contains expected word meanings of each target word and requires only minor human corrections such as the need for extra instructions to generate longer sentences. As an example, Table 8 reports our instructions for the word "ratio" with a specific sense in Latin.

Recall that our clustering approach involves two hyperparameters t^0_{sc} and t^1_{sc} to determine whether two clusters are similar enough to be merged. To tune these hyperparameters on the development sets we constructed, we first use our clustering approach to produce a prediction of label vector, denoted as $\hat{\mathcal{Y}}_w(t^0_{sc}, t^1_{sc}) = [\hat{y}_1, \ldots, \hat{y}_n]$ with $\hat{y}_i \in \{0, 1\}$ for each configuration of hyperparameters. Then, we use grid search to tune the hyperparameters based on the idea of Adjusted Mutual Information (AMI) (Vinh et al., 2009), denoted as:

$$\underset{t^0_{sc}\in(0,1), t^1_{sc}\in(0,1)}{\arg\max} \sum_{w\in W_\ell} \text{AMI}(\mathcal{Y}_w, \hat{\mathcal{Y}}_w(t^0_{sc}, t^1_{sc}))$$

where W_ℓ denotes a set of target words for each language ℓ.

A.3 Experimental Setups for Exploratory Study

Datasets. We choose a set of target word triplets that are translations in English, German and Swedish, such as {'mouse', 'Maus', 'mus'}. For each of these languages, we consider the SemEval2020 corpus specific to that language from the earlier time period (the 19 century) as the historical corpus at time $t-1$. For the present corpus at time t, we opt for a random selection of the most recent Wikipedia dump, rather than the SemEval2020 corpus from the later time period. This is because the later time periods in the three languages are substantially different, making it unreliable to compare semantic changes across languages. We provide data statistics in Table 7.

Implementation Details. Regarding the choice of our hyperparameters, we apply the same cross-lingual threshold t_{cs} to all languages, and set the threshold to the threshold t_{sc} used in the English intra-language setup, denoted as $t^{\text{All}}_{cs} = t^{\text{EN}}_{sc}$. The reason for this is the following: Since we re-align the embedding spaces of target languages (German and Swedish) to the source language English, we apply the hyperparameters that were tuned on the English development set to all languages in the inter-language setup. We set k to 14 for k-nearest neighboring words.

A.4 Analyses

Comparing Clustering Approaches. We comparably evaluate two classes of clustering approaches: (a) explicit/predetermined choice of the number clusters: K-means and Gaussian Mixture[12] and (b) implicit choice: Affinity Propagation (Frey and Dueck, 2007) and our approach. To begin, we select 8 target words in each language. Each word has two word senses with uneven frequency distribution (100:20). We use the popular metric Pu-

[12] https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html

Role	Conversations
Human	Please create sentences for the Latin word 'ratio' that has the meaning of 'reason or rationality, the ability to think logically and make sound judgments' in the format: a Latin sentence enclosed in double quotation marks, followed by an English translation enclosed in parentheses.
ChatGPT	1. "Sua ratione et prudentia, ille vir difficiles decisiones sapienter adhibebat."
	(By his reason and prudence, that man wisely made difficult decisions.)
	2. "Ratione uti poteris, neque impulsu emotionum agere debes."
	(You should use reason, not be driven by the impulse of emotions.)
	... ...(sentences examples)
Human	Good. Please create a bit longer sentences.
ChatGPT	1. "Cum homines ratione utantur, societates humanae pacem et progressum promovere possunt."
	(When people use reason, human societies can promote peace and progress.)
	2. "Ratione, quam discimus, in omni aspectu vitae uti debemus, ut optima decisiones capiamus."
	(We should use the reason we learn in every aspect of life to make the best decisions.)
	... ...(sentences examples)
Human	Good. Please write more sentences in the same format.
ChatGPT	... ...(sentences examples)

Table 8: Our instructions for creating sentences that contain the Latin word 'ratio' with a specific word sense.

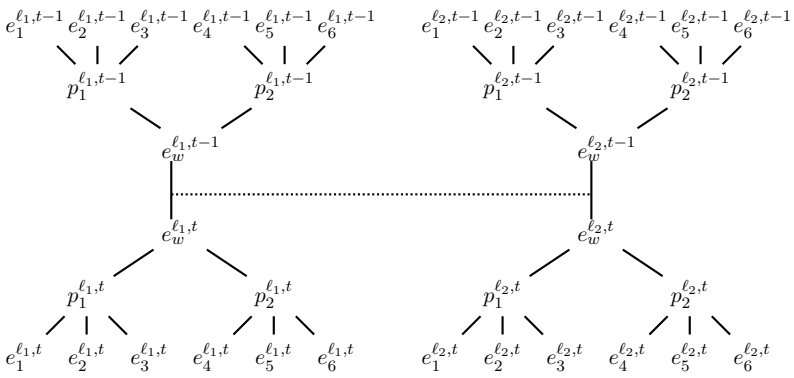


Figure 9: Representation of semantic changes of a mutual word translation pair over time in a temporal and spatial dynamic graph that links two temporal dynamic graphs in different languages.

(ℓ_1, ℓ_2)	$m(e^{\ell_1}, \mathcal{C}_w^{\ell_1})$	$m(e^{\ell_1}, \mathcal{C}_w^{\ell_2})$	$m(e^{\ell_1}, \mathcal{C}_w^{\ell_2} + b)$
(EN, DE)	0.64	0.46	0.64
(EN, SV)	0.64	0.45	0.65

Table 9: Results of embedding space alignment.

rity Scoring[13]) to evaluate clustering quality—the higher purity score indicates better quality. In Table 11, we see that our approach is quite advantageous in this setup, demonstrating its ability to capture both high and low-frequency word senses. In English, we see the performance gain of our approach is comparatively smaller. This might be because m-BERT is known to produce higher-quality embeddings in English compared to other languages, making it less susceptible to the poor quality of baseline clustering approaches.

Our Clustering Method. Figure 10 shows the relationships between the choice of thresholds (t_{sc}^0 and t_{sc}^1) and the corresponding detection accuracy. We find that the high accuracy area colored in bright yellow expands greatly as k increases, particularly for English and German. This means that the more nearest semantic neighboring words are involved, the higher detection accuracy our approach achieves. Furthermore, we see that the brightest areas across languages are shown in different locations, and these areas associate with very different configurations of t_{sc}^0 and t_{sc}^1, even for typologically similar language pairs such as English and German. This is because the SemEval2020 corpora in English and German are collected from different time

[13] https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html

	Contextualized Word Embeddings	
Components	**Previous Approaches**	**Our Approach**
Clustering Method	k-means	our clustering method
Clustering Strategy	time-independent sense clusters	time-dependent sense clusters
Semantic Representation	word embeddings	graph
Distance Metric	Euclidean distance	neighbor-based distance
Detection Criterion	frequency-based criterion	similarity between sense clusters

Table 10: Contrasting previous approaches (Kanjirangat et al., 2020; Cuba Gyllensten et al., 2020) and our approach for detecting semantic change. All these approaches combine contextualized word embeddings with a clustering method but differ in several aspects. Our neighbor-based metric is adopted in two components: our clustering method and detection criterion.

Algorithms	EN	DE	LA	SV
K-means	0.975	0.778	0.664	0.775
Gaussian Mixture	0.939	0.775	0.670	0.754
Affinity Propagation	0.891	0.741	0.686	0.662
Our Clustering	**0.994**	**0.879**	**0.877**	**0.909**

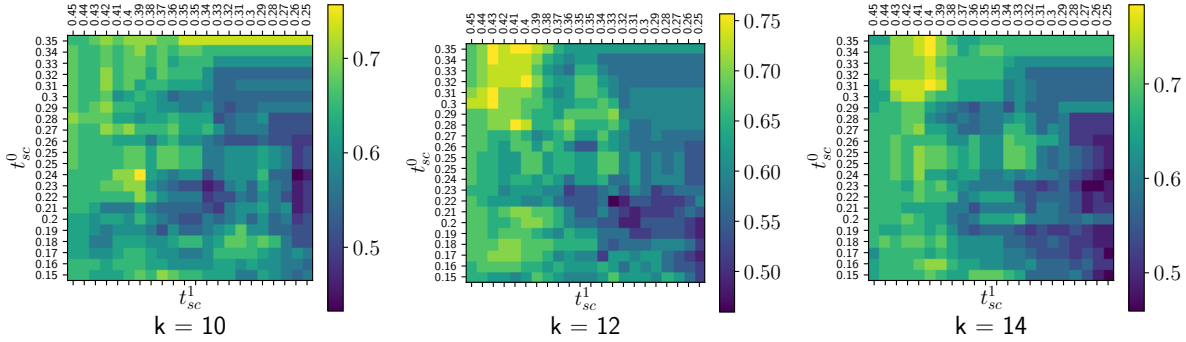
Table 11: Purity scores across four approaches in the 100:20 frequency distribution setup.

periods (see Table 4), making these two languages further apart from each other.

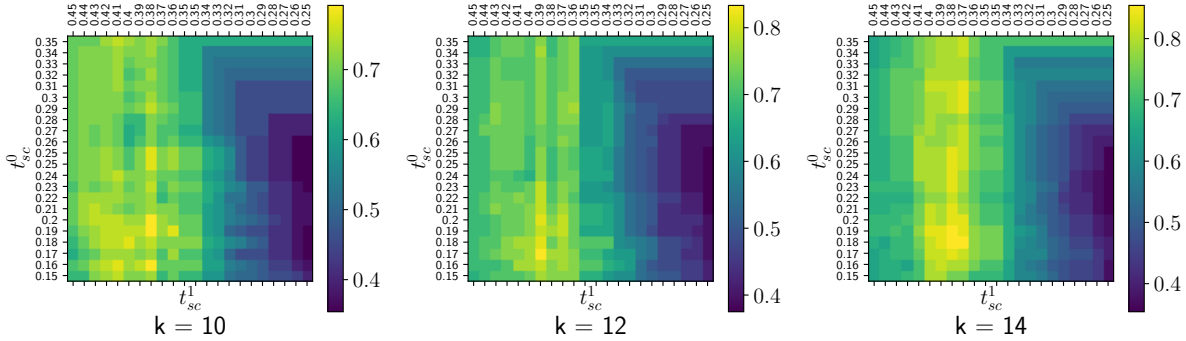
Bilingual Embedding Spaces. If two embedding spaces of languages ℓ_1 and ℓ_2 align well, they should share the same space centroid and the same topological structure. We consider the topological structure of the ℓ_1 embedding space as $m(e^{\ell_1}, \mathcal{C}_w^{\ell_1}) = \frac{1}{|\mathcal{C}_w^{\ell_1}|} \sum_{c_i \in \mathcal{C}_w^{\ell_1}} s(e^{\ell_1}, c_i)$, i.e, the average similarity between each point in $\mathcal{C}_w^{\ell_1}$ and the $\mathcal{C}_w^{\ell_1}$'s centroid e^{ℓ_1}. Therefore, $m(e^{\ell_1}, \mathcal{C}_w^{\ell_1})$ should closely match $m(e^{\ell_1}, \mathcal{C}_w^{\ell_2})$ in this case. However, Table 9 shows that the score $m(e^{\ell_1}, \mathcal{C}_w^{\ell_2})$ is much lower than $m(e^{\ell_1}, \mathcal{C}_w^{\ell_1})$, implying that the ℓ_1 and ℓ_2 embedding spaces exhibit quite different topological structures. This arises from the fact that the two embedding spaces are initially misaligned. After applying a rectified vector b, we see a close match between $m(e^{\ell_1}, \mathcal{C}_w^{\ell_1})$ and $m(e^{\ell_1}, \mathcal{C}_w^{\ell_2} + b)$ in terms of topological structure, demonstrating the effectiveness of the chosen rectification approach (Liu et al., 2020) for addressing the misalignment between embedding spaces of different languages.

A.5 Hardware Specifications and Execution Times

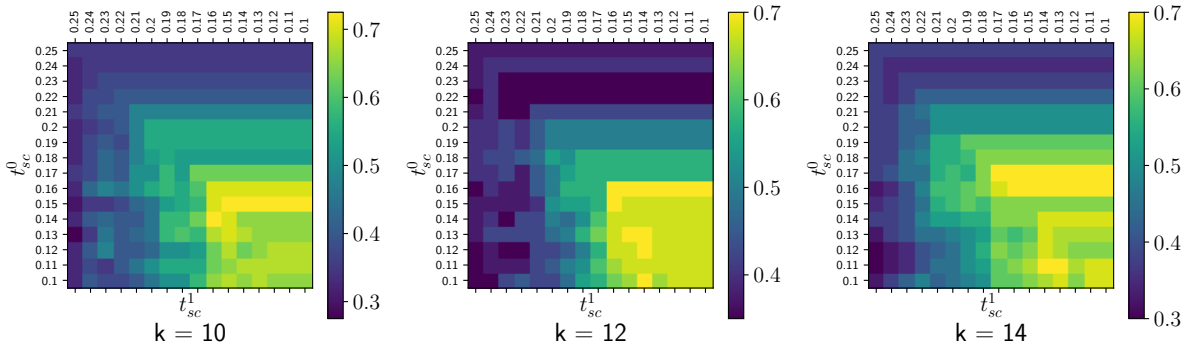
All experiments were executed on a computer featuring an AMD CPU with 8 cores, 32GB of RAM and a single RTX3060 GPU with 12GB of memory. For each target word, it takes about 60 seconds for m-BERT to generate its contextualized word embeddings within 800 sentences on GPU; our clustering method takes about 5 minutes to complete on CPU with 8 multi-processing threads.



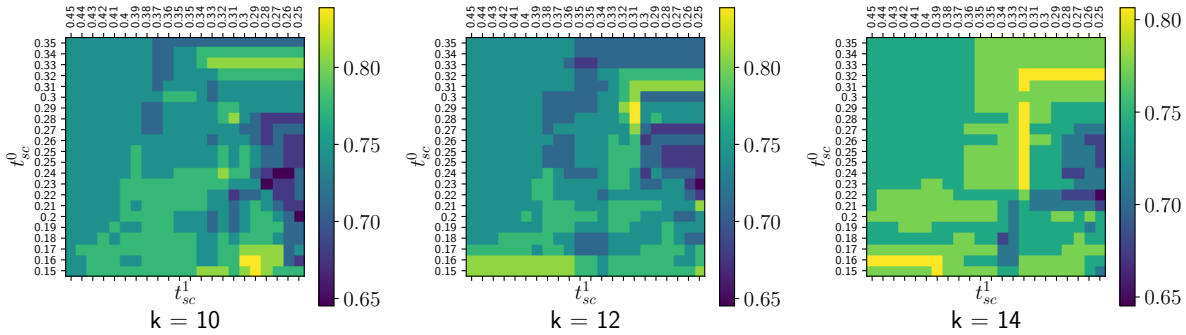
(a) English



(b) German

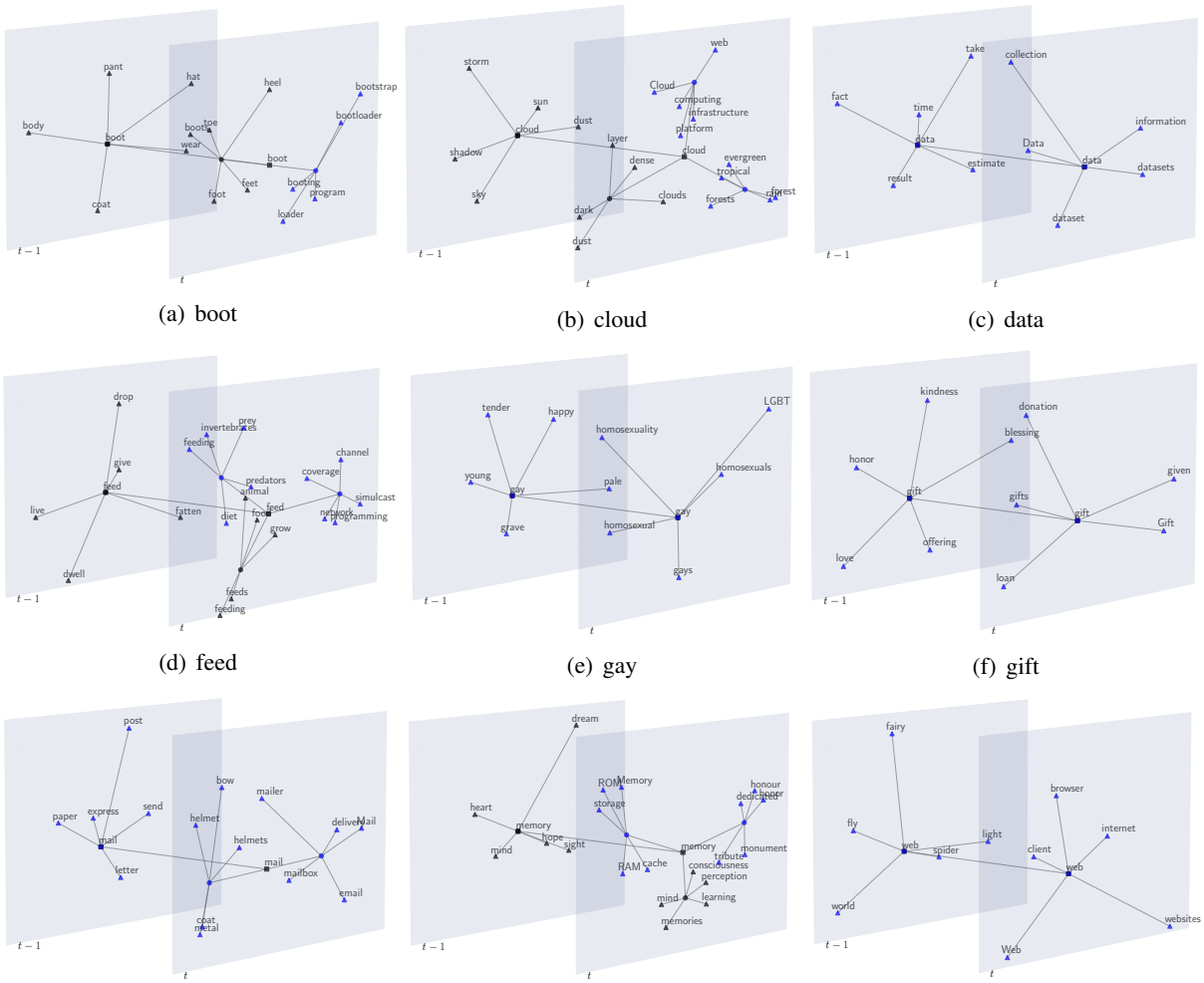


(c) Latin



(d) Swedish

Figure 10: Relationships between the threshold values and accuracies in the SemEval2020 binary classification task.

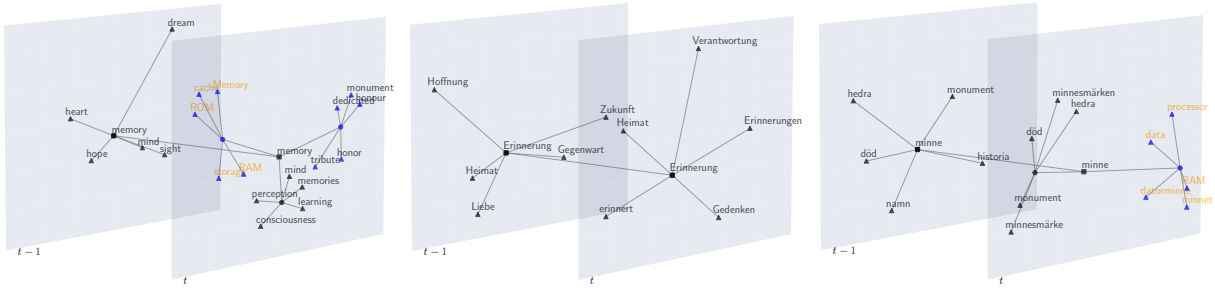


(a) boot (b) cloud (c) data

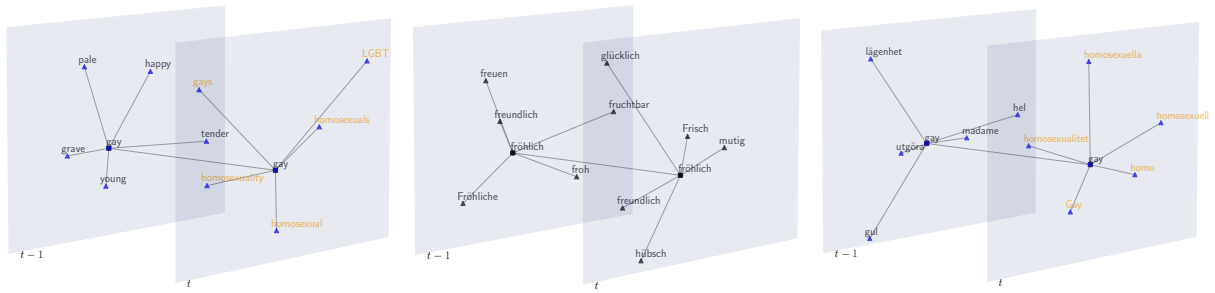
(d) feed (e) gay (f) gift

(g) mail (h) memory (i) web

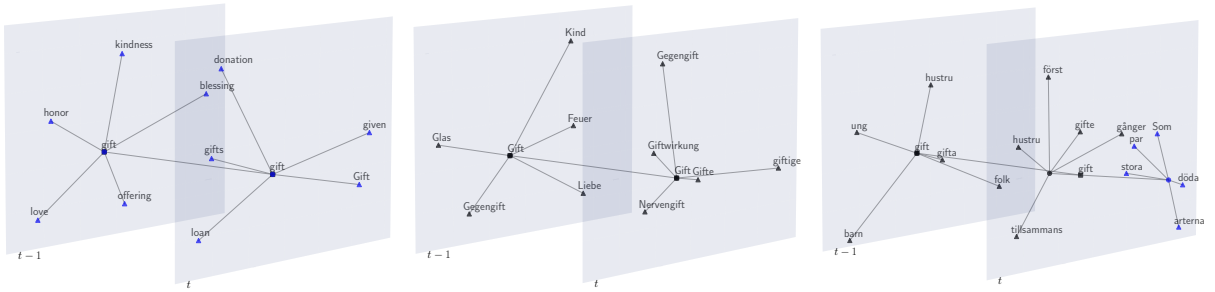
Figure 11: Representation of the detected semantic changes for the English word ‘boot’, ‘cloud’, ‘data’, ‘feed’, ‘gay’, ‘gift’, ‘mail’, ‘memory’, and ‘web’ in our temporal dynamic graph. Blue nodes at time t indicate the acquisition of new meanings, while blue nodes at time $t-1$ indicate the loss of original meanings. Black nodes indicate word meanings that remain unchanged over time.



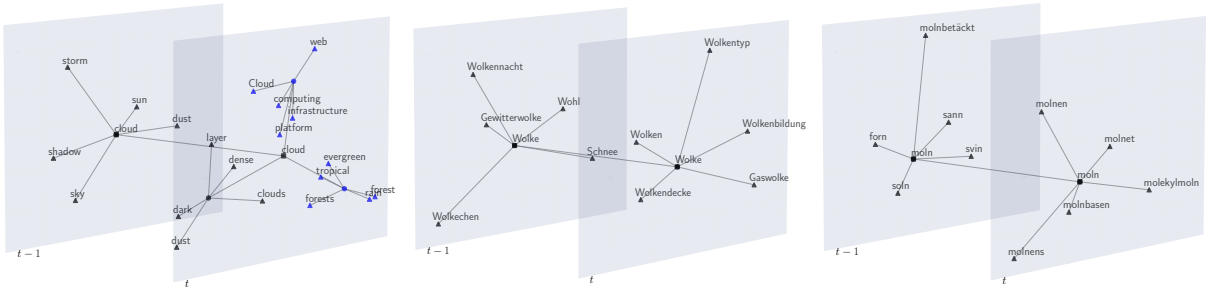
(a) {memory, Erinnerung, minne}



(b) {gay, fröhlich, gay}



(c) {gift ,Gift, gift}



(d) {cloud, Wolke, moln }

Figure 12: Representation of the detected semantic changes in word translations across English, German and Swedish, shown in our inter-language temporal dynamic graphs. Blue nodes indicate the acquisition of a new meaning at time t (the loss of an existing one at $t-1$), while black nodes indicate meanings unchanged over time. Orange text marks certain changed meanings that are considered highly similar across languages. Figure (a) {memory, Erinnerung, minne}: 'memory' and 'minne' gain the same new meaning about computer storage unit at time t while the meaning of 'Erinnerung' remains unchanged. Figure (b) {gay, fröhlich, gay}: 'gay' (English) and 'gay' (Swedish) gain the same new meaning about homosexuality at time t, and both words lose their different original meanings that they had at $t-1$. The meaning of 'fröhlich' remains unchanged. Figure (c) {gift, Gift, gift}: their semantic changes diverge greatly over time. The meaning of 'gift' (English) changes from *a notable act of giving* to *something given voluntarily without payment*. Gift (German) retains the meaning *poison* over time. 'gift' (Swedish) gains a new meaning *poison* perhaps borrowed from German. Figure (d) {cloud, Wolke, moln}: 'cloud' gains a new meaning while the meanings of 'Wolke' and 'moln' remain unchanged.