Quantifying Lexical Semantic Shift via Unbalanced Optimal Transport

Ryo Kishino[1] **Hiroaki Yamagiwa**[1]
Ryo Nagata[2,5] **Sho Yokoi**[3,4,5] **Hidetoshi Shimodaira**[1,5]
[1] Kyoto University [2] Konan University [3] NINJAL [4] Tohoku University [5] RIKEN
kishino.ryo.32s@st.kyoto-u.ac.jp, h.yamagiwa@i.kyoto-u.ac.jp,
nagata-acl2025@ml.hyogo-u.ac.jp, yokoi@ninjal.ac.jp, shimo@i.kyoto-u.ac.jp

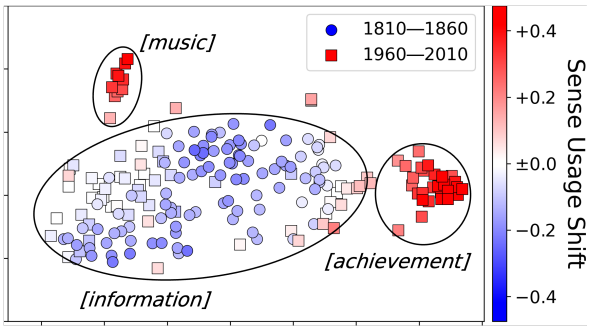
Abstract

Lexical semantic change detection aims to identify shifts in word meanings over time. While existing methods using embeddings from a diachronic corpus pair estimate the degree of change for target words, they offer limited insight into changes at the level of individual usage instances. To address this, we apply Unbalanced Optimal Transport (UOT) to sets of contextualized word embeddings, capturing semantic change through the excess and deficit in the alignment between usage instances. In particular, we propose Sense Usage Shift (SUS), a measure that quantifies changes in the usage frequency of a word sense at each usage instance. By leveraging SUS, we demonstrate that several challenges in semantic change detection can be addressed in a unified manner, including quantifying instance-level semantic change and word-level tasks such as measuring the magnitude of semantic change and the broadening or narrowing of meaning.

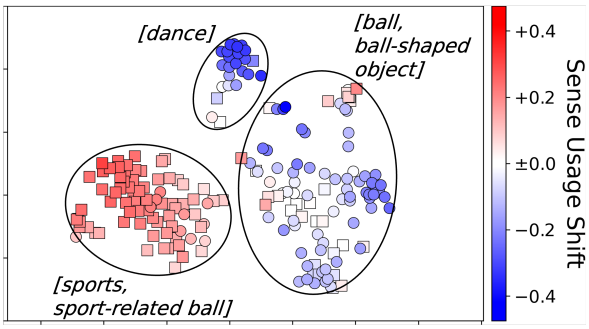
1 Introduction

Lexical semantic change detection is the task of identifying words that change their meaning over time, as well as determining which specific senses have disappeared or emerged (Periti and Montanelli, 2024). Recently, methods leveraging word embeddings from a diachronic corpus pair have been studied. Previous approaches detect semantic change by measuring differences in contextualized embeddings of a target word across two corpora (Giulianelli et al., 2020; Montariol et al., 2021; Aida and Bollegala, 2023).

These methods focus on the entire set of usage instances. From a linguistic point of view, however, it is critical to focus on individual instances as well as on the overall differences. For example, it is important to quantify the extent to which a disappearing or emerging sense of a word loses or acquires its popularity compared to the other senses in order to reveal how and why the word has undergone the semantic change. Unfortunately, there have been almost no studies on this topic.



(a) *record*



(b) *ball*

Figure 1: t-SNE visualization of the contextualized embeddings for each usage instance of a target word in the diachronic corpus pair. The color of each instance represents its Sense Usage Shift (SUS). The relative frequency of the target word usage in the senses of the red instances has increased compared to its usage in the other senses, while that of the blue instances has decreased. (a) For the word *record*, instances in the sense of *[music]* and *[achievement]*, whose usage frequencies have increased, exhibit high SUS values. (b) For the word *ball*, instances in the sense of *[dance]*, whose usage frequency has decreased, exhibit low SUS values. For more details, refer to Section 4.3, and for additional word examples, see Appendix A.

Our code is available at https://github.com/ryo-lyo/Semantic-Shift-via-UOT.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15913–15933

	Corpus	Instance	Sense	SUS
Top 3	1960–2010	... So did Sire ***Records***...	undefined (*[music]*)	0.47
	1960–2010	... a team with the third-worst ***record***...	*[achievement]*	0.45
	1960–2010	... the AMCU single-season ***record***...	*[achievement]*	0.45
Bottom 3	1810–1860	... interpretations of the Mosaic ***record***...	*[information]*	-0.23
	1810–1860	... the ***records*** of a professed revelation...	*[information]*	-0.24
	1810–1860	... the ***record*** of whose wisdom is included in...	*[information]*	-0.25

Table 1: The top 3 and bottom 3 usage instances of the target word *record* based on SUS values. Instances with high SUS values correspond to senses whose usage has increased across the diachronic corpus pair, while instances with low SUS values correspond to senses whose usage has decreased, as discussed in Section 4.3. For other target words, see Table 5 in Appendix A. For more details, including 'undefined', refer to Appendix D.

To address this limitation, we apply Unbalanced Optimal Transport (UOT) between the two sets of contextualized embeddings of a target word obtained from a diachronic corpus pair, focusing on the excess and deficit in the alignment between usage instances. Using this alignment discrepancy, we propose a novel measure called Sense Usage Shift (SUS), which quantifies how much the relative frequency of the word usage in a word sense of each instance has changed across the corpora.

Fig. 1 illustrates that word usage in the word sense of instances with high SUS values has become frequent in the modern corpus, whereas usage in the sense of instances with low SUS values has decreased. Table 1 shows several usage instances with high or low SUS values. Although the previous work (Montariol et al., 2021) has applied standard Optimal Transport (OT) for detecting semantic change, the balanced alignment fails to capture information about individual instances, as illustrated in Fig. 2 and Fig. 3.

Moreover, by leveraging the SUS values calculated for each usage instance, we address various tasks in semantic change in a unified manner. Specifically, for a given target word, SUS enables (1) the quantification of semantic change at the instance level, (2) the quantification of semantic change at the word level, and (3) the quantification of the extent to which the meaning of a target word has broadened or narrowed. Experiments demonstrate that the proposed SUS-based methods achieve performance comparable to or better than existing approaches for these tasks.

2 Related Work

2.1 Lexical semantic change detection

One of the major approaches to lexical semantic change detection is to align two sets of static word embeddings to measure semantic differences between a diachronic corpus pair. This can be done either by assuming linear transformations (Kulkarni et al., 2015; Hamilton et al., 2016) or without the linear assumption (Yao et al., 2018; Aida et al., 2021) between the two embedding spaces. However, static embedding-based methods can only handle word-level semantic change, providing no information on changes at the sense level or the individual instance level.

To address this, methods leveraging contextualized word embeddings for sets of usage instances of a target word, obtained from old and modern corpora using language models such as BERT (Devlin et al., 2019), have been proposed. These methods are broadly categorized into sense-based and form-based approaches (Giulianelli et al., 2020; Periti and Montanelli, 2024) as follows.

Sense-based approach. This approach aims to estimate, either directly or indirectly, the frequency of word usage corresponding to specific senses of the target word in two given corpora. The estimated sense counts are then directly used to compute the degree of semantic change. Giulianelli et al. (2020); Rother et al. (2020); Kutuzov and Giulianelli (2020); Periti et al. (2022) applied clustering algorithms to two sets of contextualized embeddings of the target word to identify sense counts. Clustering methods such as K-means and Affinity Propagation (Frey and Dueck, 2007) are commonly used for this purpose. However, K-means faces the non-trivial challenge of determining the number of senses, while Affinity Propagation suffers from unstable clustering results (Periti and Montanelli, 2024).

Form-based approach. This approach detects semantic change by comparing the probability distributions of contextualized word embeddings from the two corpora. Aida and Bollegala (2023) and

Nagata et al. (2023) assume that embeddings follow a normal distribution and a von Mises-Fisher (vMF) distribution, respectively, and define the degree of semantic change as the difference between the two distributions.

Unlike the sense-based approach, the form-based approach bypasses sense identification, which is likely to result in more stable semantic change detection. As mentioned in Eqs. (3) and (8) of Nagata et al. (2023), it is also possible to capture semantic change at the instance level by using the ratio of probability densities in the two corpora. While Nagata et al. (2023) discuss this approach, to the best of our knowledge, prior research has not quantitatively evaluated the detection of semantic change at the instance level.

2.2 Comparing distributions of embeddings with optimal transport

OT is a method for measuring the distance between two probability distributions through complete alignment, while UOT allows for excess and deficit in the alignment. In natural language processing, OT is widely used to compute the distance between two documents based on embeddings (Kusner et al., 2015; Yokoi et al., 2020), and UOT is also utilized for this purpose (Wang et al., 2020; Chen et al., 2020; Swanson et al., 2020; Arase et al., 2023; Zhao et al., 2020).

In semantic change detection, Montariol et al. (2021) were the first to utilize OT to compute the degree of semantic change for a target word. Their approach involved clustering the contextualized embeddings of a target word obtained from two corpora and applying OT to the sets of centroids corresponding to each sense. Consequently, unlike the proposed method, this approach does not provide information about individual usage instances except through the identified sense clusters. On the other hand, Pranjic et al. (2024) directly applied OT to two sets of contextualized embeddings. However, to the best of our knowledge, no existing studies have utilized UOT for semantic change detection, as proposed in our method.

3 Background on Optimal Transport

3.1 Problem setting

We aim to identify how the meanings of a target word w change across a diachronic corpus pair. A context containing the target word w is referred to as *a usage instance of* w. For simplicity, it is referred to as *an instance*. The sense of w in a usage instance is referred to as *the sense of an instance*. Let the set of m usage instances of the target word w belonging to the old corpus be $\{s_i^w\}_{i=1}^m$. Likewise, let the set of n usage instances of w belonging to the modern corpus be $\{t_j^w\}_{j=1}^n$. Using a pre-trained language model, the contextualized embeddings $\boldsymbol{u}_i^w$ and $\boldsymbol{v}_j^w \in \mathbb{R}^d$ are calculated for each usage instance s_i^w and t_j^w, respectively. Hereafter, the superscript w is omitted for brevity. To detect how the meaning of w has changed between the corpora, we measure the difference between the distributions $\{\boldsymbol{u}_i\}_{i=1}^m$ and $\{\boldsymbol{v}_j\}_{j=1}^n$.

3.2 Optimal transport

For $n \in \mathbb{N}$, let $\mathbf{1}_n$ denote an n-dimensional vector with all elements equal to 1 and $\mathbb{R}_+ = [0, \infty)$ represent the set of non-negative real numbers.

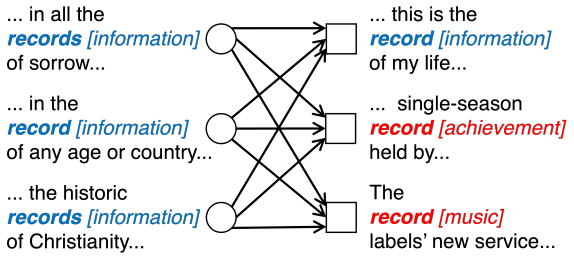
Let $a_i \in \mathbb{R}_+$ and $b_j \in \mathbb{R}_+$ denote the weights associated with the contextualized word embeddings u_i and v_j, respectively. satisfying $\sum_i a_i = \sum_j b_j = 1$. Thus, these weighted embeddings can be regarded as probability distributions. Let $C_{ij} \in \mathbb{R}_+$ denote the transportation cost between the embeddings $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$, and $T_{ij} \in \mathbb{R}_+$ the amount of transportation. The OT problem, which minimizes the total transportation cost, is formulated as follows:

$$\begin{aligned} &\min_{\boldsymbol{T} \in \mathbb{R}_+^{m \times n}} \sum_{i,j} T_{ij} C_{ij} \\ &\text{s.t. } \boldsymbol{T}\mathbf{1}_n = \boldsymbol{a}, \quad \boldsymbol{T}^\top \mathbf{1}_m = \boldsymbol{b}, \end{aligned} \tag{1}$$

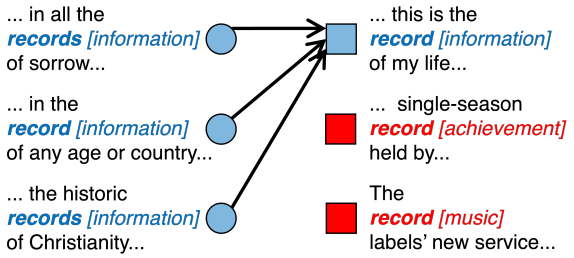
where $\boldsymbol{a} = (a_1, \ldots, a_m)^\top$, $\boldsymbol{b} = (b_1, \ldots, b_n)^\top$, and $\boldsymbol{T} = (T_{ij})$. The optimization variable $\boldsymbol{T}$ is referred to as the transportation matrix, which can be interpreted as providing the alignment between the two sets of the embeddings. Furthermore, the coupling constraints $\boldsymbol{T}\mathbf{1}_n = \boldsymbol{a}, \boldsymbol{T}^\top \mathbf{1}_m = \boldsymbol{b}$ ensure that there is no excess or deficit in the alignment between the two sets of the embeddings.

3.3 Limitation of OT in semantic change detection

Pranjic et al. (2024) utilized the value of the total transportation cost $\sum_{i,j} T_{ij} C_{ij}$ with the optimized $\boldsymbol{T}$ as a measure of semantic change at the word level. Standard OT establishes balanced alignment between embeddings (Fig. 2a). However, when w undergoes semantic change, substantial shifts can occur between the two corpora, such as the emergence of new senses or the disappearance of



(a) Balanced alignment in OT



(b) Unbalanced alignment in UOT

Figure 2: Illustration of the difference between OT and UOT. Circles and squares represent usage instances of the word *record* in the old and modern corpora, respectively, and the arrows indicate transportation. In the example above, the two bottom squares on the modern side represent instances in the newly emerged senses. (a) OT enforces a balanced alignment between the two sets of instances, which fails to model semantic change, as discussed in Section 3.3. (b) UOT allows for excess or deficit in alignment. The lack of transported weight from the old side to the two bottom squares on the modern side reflects the increased word usage of these new senses in the modern corpus, as described in Section 4.1. See Fig. 3 for a visualization of the transport matrix based on actual numerical experiments.

existing ones. These shifts pose challenges for OT, as its balanced alignment assumption may not fully account for such large semantic changes. As can be seen from the example of the actual transportation matrix $\boldsymbol{T}$ shown in Fig. 3, OT frequently conducts transportation between different senses. As a result, while OT can capture overall semantic change at the word level, it may not fully capture semantic changes in individual usage instances.

4 Proposed Method

4.1 Modeling semantic change with UOT

The balanced alignment in OT, enforced by the coupling constraints in (1), does not always reflect real-world semantic changes, especially in cases where new senses emerge or existing senses disappear. To address this issue, we consider the following

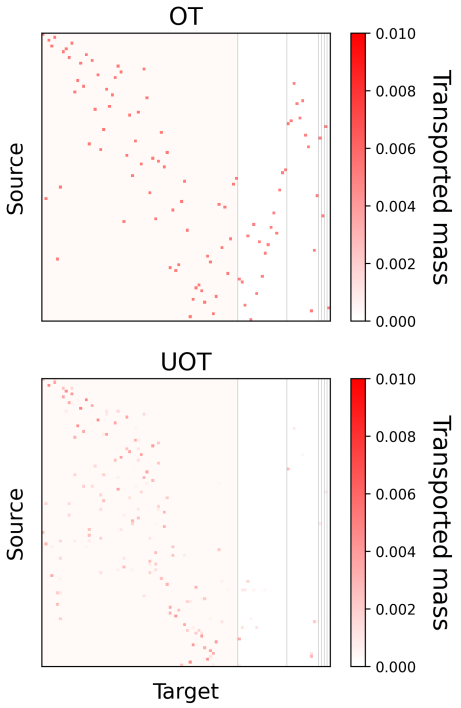


Figure 3: The transportation matrix of OT and UOT for the target word *record*. Red-shaded areas indicate alignments between instances in the same word sense across the two corpora, while white areas indicate alignments between different senses. (Left) OT conducts transportation across different senses. (Right) UOT reduces such transportation by allowing alignment discrepancy. For additional word examples, refer to Fig. 8 in Appendix A.

problem, which relaxes these constraints in (1):

$$\min_{\boldsymbol{T}\in\mathbb{R}_+^{m\times n}} \sum_{i,j} T_{ij}C_{ij} + \lambda_1 D_1(\boldsymbol{T}\mathbf{1}_n, \boldsymbol{a}) + \lambda_2 D_2(\boldsymbol{T}^\top \mathbf{1}_m, \boldsymbol{b}). \tag{2}$$

Here, $D_1 : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ and $D_2 : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ are penalty functions, and λ_1, λ_2 are hyperparameters that control the degree of penalization. The problem (2) is referred to as Unbalanced Optimal Transport (UOT). UOT allows for excess or deficit in the alignment by incurring a certain cost. In our experiments, we assigned uniform weights to each embedding and employed the L2 error as the penalty function. For details, see Section 5.

We focus on the alignment discrepancy represented by $\boldsymbol{T}$ in (2) (Fig. 2b). As can be seen from the example in Fig. 3, the transportation in UOT tends to be restricted within the same sense, allowing the alignment excess or deficit to reflect changes in the usage frequency of each sense. With respect to s_i, if less weight is transported to the modern corpus than the original weight a_i, this in-

dicates that instances in the word sense of s_i are relatively scarce in the modern corpus compared to other w senses. In other words, the usage of w in the word sense of s_i has decreased. Similarly, for t_j, if less weight is received from the old corpus than the original weight b_j, this suggests that instances in the word sense of t_j are relatively scarce in the old corpus. In other words, the usage of w in the word sense of t_j has increased.

4.2 Sense usage shift

To quantitatively measure semantic change via the excess or deficit in the alignment from UOT, we define Sense Usage Shift (SUS) as a measure of how the frequency of word usage in the word sense of each usage instance s_i or t_j has changed relative to the usage frequencies in the other senses. SUS is defined as follows:

$$\mathrm{SUS}(s_i) = -(a_i - \sum_j T_{ij})/a_i, \quad (3)$$

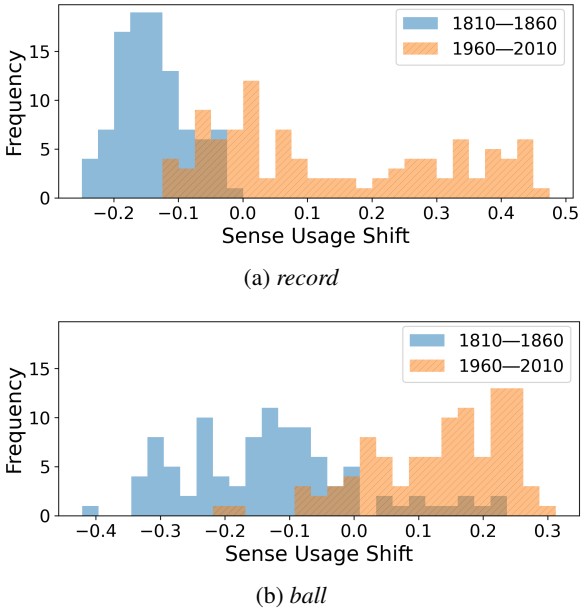
$$\mathrm{SUS}(t_j) = (b_j - \sum_i T_{ij})/b_j. \quad (4)$$

In other words, SUS represents the excess or deficit in the alignment normalized by the original weight. From the discussion in Section 4.1, a higher SUS value for a usage instance indicates more frequent usage of the target word in the word sense of that instance in the modern corpus. Conversely, a lower SUS value suggests less frequent usage of the target word in that sense. In particular, in (3), the sign is inverted to ensure that a positive SUS value for an instance corresponds to increased usage in its sense.

4.3 Understanding semantic change using SUS

To gain deeper insights into SUS defined in (3) and (4), we visualize SUS values. We focus on the target words *record* and *ball*, each with 100 usage instances in the Diachronic Word Usage Graph (DWUG) dataset (Schlechtweg et al., 2021, 2024). In this dataset, the meaning of *record* has broadened, while that of *ball* has narrowed across the two corpora. Using a pre-trained language model, we compute the contextualized word embeddings $\{\boldsymbol{u}_i\}_{i=1}^{100}$ and $\{\boldsymbol{v}_j\}_{j=1}^{100}$. For more details on the experimental setup, refer to Section 5.

Visualizing semantic change with SUS. Fig. 1 presents a two-dimensional visualization of the contextualized embeddings of the target words *record* and *ball*, generated using t-SNE[1]. In this plot, the clusters[2] represent the senses of the target words. We analyzed the usage instances within each cluster and identified the sense associated with them. Additional visualization examples for other target words are provided in Fig. 7a, and usage instances with high or low SUS values are listed in Table 5 in Appendix A.



(a) *record*

(b) *ball*

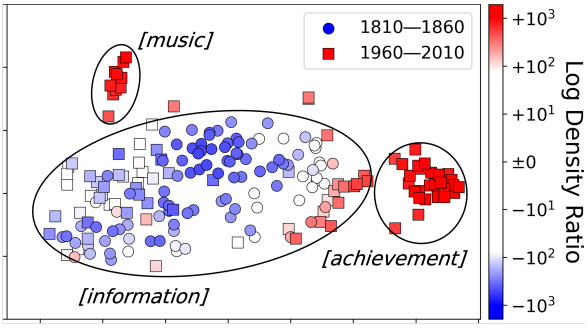
Figure 4: The distribution of SUS values for usage instances of the target words. (a) For the target word *record*, the variance of SUS values is larger in the modern corpus. This indicates that some usage instances associated with senses in the old corpus also appear in the modern corpus. Therefore, the SUS distribution suggests that *record* is used with a broadened meaning in the modern corpus. (b) For the target word *ball*, conversely, the variance of SUS values is larger in the old corpus, indicating that *ball* is used with a narrowed meaning in the modern corpus.

By examining the clusters in these figures, it can be observed that usage instances associated with senses whose frequency has increased or newly emerged have high SUS values, while those associated with senses whose frequency has decreased or disappeared have low SUS values.

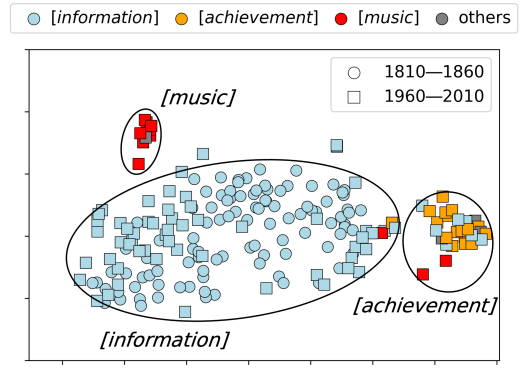
Distribution of SUS. Furthermore, the distribution of SUS values allows us to interpret whether

[1] In the t-SNE configuration, we set the perplexity to 30, the default value in the scikit-learn implementation.

[2] In Fig. 1, the clusters and their sense labels were manually examined by the authors to represent the sense of the instances. While not all instances were explicitly or strictly annotated, efforts were made to maintain consistency within each cluster. The same procedure was applied to Fig. 7 in Appendix A.



(a) LDR



(b) DWUG

Figure 5: Re-rendering of the t-SNE visualization for the word *record* in Fig. 1a. The color of each instance represents (a) the Log-Density Ratio (LDR) and (b) the gold sense in the DWUG dataset. For details and additional word examples with LDR coloring, refer to Fig. 7b in Appendix A. For details on the gold sense, see Section 5.2.

the meaning of a word has broadened or narrowed. Fig. 4 illustrates the distribution of SUS values for usage instances of the target words. If the SUS values are regarded as surrogate values for embeddings, a larger variance in SUS values indicates a broader semantic scope. Accordingly, we infer that the meaning of *record* has broadened, whereas the meaning of *ball* has narrowed.

4.4 Log-density ratio

Definition and computation. As a baseline for SUS in the form-based approach, we consider the log-density ratio (LDR). LDR is defined as the logarithm of the ratio of two probability density functions, which are parametrically estimated from the embeddings in the old and modern corpora. The estimation of density ratios has been extensively studied in statistics and machine learning (Sugiyama et al., 2012).

In the context of semantic change detection, LDR is mentioned in Eq. (8) of Nagata et al. (2023). Note that they use the term log-likelihood ratio to refer to the LDR at a single instance, although it typically refers to comparisons between densities over the entire dataset.

In our setup, we compute LDR as follows. First, following Nagata et al. (2023), we assume that the normalized old and modern embeddings follow a distinct von Mises-Fisher (vMF) distribution[3], p_S and p_T. Then, we perform maximum likelihood estimation of the parameters using an approximate closed-form solution, as described in Appendix B. Using the estimated parameters, we compute

$$\mathrm{LDR}(s_i) = \log \frac{p_T(\tilde{\boldsymbol{u}}_i)}{p_S(\tilde{\boldsymbol{u}}_i)},$$

$$\mathrm{LDR}(t_j) = \log \frac{p_T(\tilde{\boldsymbol{v}}_j)}{p_S(\tilde{\boldsymbol{v}}_j)},$$

where $\tilde{\boldsymbol{u}}_i = \boldsymbol{u}_i / \|\boldsymbol{u}_i\|$ and $\tilde{\boldsymbol{v}}_j = \boldsymbol{v}_j / \|\boldsymbol{v}_j\|$.

Comparison between SUS and LDR. A visual comparison of Fig. 1 and Fig. 5a suggests that both SUS and LDR roughly reflect sense clusters appropriately. However, referring to the gold senses[4] provided by the DWUG dataset in Fig. 5b, it is evident that SUS reflects the clusters more accurately than LDR. In particular, in the rightmost part of the cluster corresponding to the sense *[information]*, there is a red region where circles and boxes are mixed. LDR tends to separate this region as a distinct cluster, although this separation does not match the gold senses. SUS, on the other hand, better preserves the overall cluster structure. For preprocessing on color configuration, see Appendix A.

LDR values often fluctuate substantially, making their estimation in high-dimensional spaces unstable. SUS offers a non-parametric alternative that bypasses this instability via UOT.

A detailed comparison between SUS and LDR in semantic change detection tasks will be presented later in Sections 6 and 7.

4.5 Aggregating SUS to quantify word-level semantic change

SUS quantifies semantic change at the level of individual usage instances. By aggregating SUS across all usage instances of a target word, we quantify the degree of word-level semantic change.

[3]The probability density function of the vMF distribution with mean direction parameter $\boldsymbol{\mu}$ and concentration parameter κ is given by $p(\boldsymbol{x} \mid \boldsymbol{\mu}, \kappa) \propto \exp(\kappa \boldsymbol{\mu}^\top \boldsymbol{x})$, where $\|\boldsymbol{x}\| = 1$.

[4]In the dataset used in the experiments, word senses are defined automatically based on human annotations. See Section 5.2 for further details.

We consider two aspects of word-level semantic change. The first is unsigned change, which focuses on the magnitude of semantic shift, regardless of its direction. The second is signed change in the scope of word meaning, which captures the direction of the change. We propose metrics to quantify the degree of each change, which are evaluated in Section 7. Due to space constraints, only one metric for each aspect of change is presented in the main text, while additional metrics are provided in Appendix G.

Unsigned change. We define the metric f_{SUS} to quantify the magnitude of semantic shift. It is computed as the absolute difference between the mean values of the SUS distributions for the two corpora. This metric is given by:

$$f_{\mathrm{SUS}}(w) = \left| \frac{1}{m} \sum_{i=1}^{m} \mathrm{SUS}(s_i) - \frac{1}{n} \sum_{j=1}^{n} \mathrm{SUS}(t_j) \right|.$$

Signed change. For a target word, let the variances of the SUS distributions in the old and modern corpora be $V_S = \mathrm{Var}(\{\mathrm{SUS}(s_i)\}_{i=1}^{m})$ and $V_T = \mathrm{Var}(\{\mathrm{SUS}(t_j)\}_{j=1}^{n})$, respectively. As discussed in Section 4.3, the difference in SUS variances between the two corpora reflects the broadening or narrowing of word meaning. Based on this observation, we define the metric g_{SUS}, which quantifies the change in semantic scope, as:

$$g_{\mathrm{SUS}}(w) = \log \frac{V_T}{V_S}.$$

5 Experimental Setup

5.1 Experimental configuration

Dataset. In the experiments, we used the Diachronic Word Usage Graph (DWUG) dataset for English version 3 (Schlechtweg et al., 2021, 2024). It contains 46 target words. For each target word, about 100 usage instances are provided for an old time period (1810–1860) and also for a new time period (1960–2010). For details on DWUG, refer to Appendix D. In particular, an overview of the target word *record* is provided in Table 8 therein.

In addition, to verify that the proposed method does not overfit to a single dataset, we also conducted experiments on DWUG ES (Zamora-Reina et al., 2022), a dataset designed to detect semantic change in Spanish. This dataset was constructed following the same procedure used to construct English DWUG dataset. The results on this dataset are presented in Appendix H.

Embedding extraction. XL-LEXEME (Cassotti et al., 2023) was used to obtain contextualized word embeddings. The dimension of embeddings is $d = 1024$. See Appendix E for details on the process of XL-LEXEME calculating the contextualized embeddings using the DWUG dataset.

UOT parameters. We set uniform weights for each usage instance: $\boldsymbol{a} = \left(\frac{1}{m}, \ldots, \frac{1}{m}\right)^\top$, $\boldsymbol{b} = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)^\top$. We defined the transportation cost between instances using the cosine distance: $C_{ij} = 1 - \cos(\boldsymbol{u}_i, \boldsymbol{v}_j)$. For penalizing excess or deficit in the alignment, we used the L2 error: $D_1(\boldsymbol{T}\mathbf{1}_n, \boldsymbol{a}) = \|\boldsymbol{T}\mathbf{1}_n - \boldsymbol{a}\|_2^2$, $D_2(\boldsymbol{T}^\top\mathbf{1}_m, \boldsymbol{b}) = \|\boldsymbol{T}^\top\mathbf{1}_m - \boldsymbol{b}\|_2^2$. The majorization-minimization (MM) algorithm provided by Python Optimal Transport (Flamary et al., 2021) was used for implementation. Following previous work on UOT (Chapel et al., 2021), we set $\lambda_1 = \lambda_2 = \lambda$. The value of λ was kept consistent across all target words. For qualitative analysis in Section 4.3, λ was fixed at $\lambda = 100$. For quantitative evaluation using SUS in Sections 6 and 7, λ was determined for each experiment by using a validation set.

Regarding the weights $\boldsymbol{a}$ and $\boldsymbol{b}$, one alternative is to use the norms of word embeddings, as proposed by Yokoi et al. (2020). However, this choice had little impact on performance in the task described in Section 7. This is likely because all embeddings represent the same target word, and their norms do not vary substantially.

As a penalty function, the Kullback-Leibler divergence has also been used in UOT (Chapel et al., 2021; Arase et al., 2023), but we consider that using the L2 error is more natural for focusing on the excess and deficit of alignment.

5.2 Sense frequency distribution (SFD)

Gold SFD. Here, we describe the gold standard[5] used to evaluate the performance of the proposed methods and the baselines in the experiments. In the DWUG dataset, based on human annotations, each usage instance of a target word is computationally assigned a sense $k = 1, \ldots, K^*$. Note that the senses themselves are not human-annotated, as described in Appendix D. In this study, we regard these sense labels in the DWUG dataset as the gold senses. Since the sense inventory is the same in the old and modern corpora, we can define the sense frequency distribution (SFD), representing

[5]The asterisk ($*$) in mathematical notation used in the experiments indicates the gold standard.

Method	Approach	Instance-level	Sense-level
τ_{SUS}	SUS-based	**0.46**	0.83
τ_{LDR}	form-based	0.40	0.70
τ_{WiDiD}	sense-based	0.31	**0.84**

Table 2: Performance of methods for predicting the instance-level and sense-level change score.

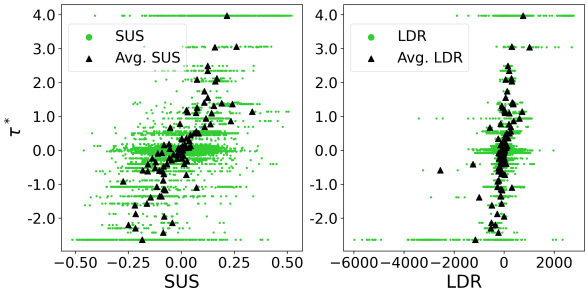


Figure 6: The relationship between the gold change score τ^* derived from DWUG and the values of SUS (left) and LDR (right) for *all* instances. The average values for each τ^* are marked with ▲. The Spearman rank correlation computed over ● corresponds to instance-level performance and ▲ to sense-level performance.

the frequency of each sense k in both corpora. Let X_k^* and Y_k^* denote the total number of instances associated with sense k in the old and modern corpora, respectively. The SFDs are then expressed as $X^* = (X_1^*, \ldots, X_{K^*}^*)$ and $Y^* = (Y_1^*, \ldots, Y_{K^*}^*)$. Although the SFDs derived from DWUG may not always reflect the true SFDs, we assume them to be sufficiently reliable for our analysis.

Baseline SFD. We used WiDiD (Periti et al., 2022), a sense-based method, as the baseline for evaluation experiments in Sections 6 and 7. Periti and Tahmasebi (2024) showed that among sense-based approaches, this method achieves SOTA performance in the task described in Section 7.1. WiDiD clusters the embeddings of the target word w from the old and modern corpora, $\{\boldsymbol{u}_i\}_{i=1}^m \cup \{\boldsymbol{v}_j\}_{j=1}^n$, deriving the SFDs $\hat{X} = (\hat{X}_1, \ldots, \hat{X}_{\hat{K}})$ and $\hat{Y} = (\hat{Y}_1, \ldots, \hat{Y}_{\hat{K}})$ in each corpus.

6 Evaluation of Quantifying Instance-Level Semantic Change

In this section, we investigate whether SUS is effective for the task of detecting semantic change at the individual usage instance level through numerical experiments. By using SUS, we identify the degree of change in the frequency of the word usage associated with each sense. For details of all experiments, refer to Appendix F.

The gold change score for performance evaluation and the baseline change score for comparison are constructed using a sense-based approach with SFD. While the sense-based approach can only measure semantic change at the sense level, rather than at the individual usage instance level, it serves as a proxy in the experiments.

Given two SFDs X and Y, we define the change score[6] for the instance s_i, or t_j, as $\tau(s_i; X, Y) = \log \frac{Y_k}{X_k}$ where k is the sense identified for s_i. The gold change score for a usage instance of a target word is defined as $\tau^*(\cdot) = \tau(\cdot; X^*, Y^*)$. As a baseline, using the SFDs $\hat{X}$ and $\hat{Y}$ estimated by the sense-based approach WiDiD, we calculated an estimate of τ^* as $\tau_{\mathrm{WiDiD}}(\cdot) = \tau(\cdot; \hat{X}, \hat{Y})$.

To estimate τ^*, our proposed method directly uses the SUS value as $\tau_{\mathrm{SUS}}(\cdot) = \mathrm{SUS}(\cdot)$. As a direct baseline for SUS, we also use the LDR value as $\tau_{\mathrm{LDR}}(\cdot) = \mathrm{LDR}(\cdot)$.

Table 2 shows the Spearman rank correlation between the gold change scores τ^* and the change score produced by each method. To clarify the situation, see Fig. 6, which shows the values of SUS and LDR for all usage instances of the target words on the horizontal axis and their corresponding gold change scores τ^* on the vertical axis. SUS outperforms the other methods, providing more accurate predictions of changes in word usage frequency for the sense of each instance. Since instances with the same sense share the same τ^*, we compute the Spearman rank correlation between the gold change scores and the average SUS values for instances with the same τ^*. We call this sense-level semantic change. SUS achieves a Spearman rank correlation of 0.83, indicating its effectiveness in capturing changes in usage frequency. In contrast, LDR shows a lower correlation of 0.70, demonstrating its inferior performance. While WiDID performs well for sense-level semantic change detection, it is worth noting that the gold scores are based on sense-level semantic change, which inherently favors WiDID as a metric.

7 Evaluation of Quantifying Word-Level Semantic Change

We evaluate the validity of the SUS-based metrics f_{SUS} and g_{SUS} for quantifying semantic change at the word level, as defined in Section 4.5. Specifically, we examine the correlation between these

[6] In cases where $X_k = 0$ or $Y_k = 0$, the values are imputed using the minimum or maximum change scores across all words, respectively.

Method	Approach	Spearman
f_{SUS}	SUS-based	0.69
f_{OT}	form-based	**0.71**
f_{APD}	form-based	**0.71**
f_{LDR}	form-based	0.31
f_{WiDiD}	sense-based	0.45
f_{APDP}	sense-based	0.51

Table 3: Performance of methods for measuring the magnitude of word-level semantic change. For the complete results, refer to Table 11 in Appendix G.1.

metrics and the gold change scores derived from the gold SFDs in DWUG.

In this section, we normalize the SFDs X and Y of w from the old and modern corpora and denote them as P and Q, respectively. For two probability distributions P and Q, we define the metrics f and g to quantify unsigned and signed change, respectively, as follows:

$$f(P,Q) = \mathrm{JSD}(P,Q),$$
$$g(P,Q) = H(Q) - H(P).$$

Here, $\mathrm{JSD}(P,Q)$ denotes the Jensen-Shannon divergence between P and Q, while H denotes the entropy, which represents the spread of the distribution; thus, the entropy difference quantifies the change in semantic scope.

7.1 Quantifying the magnitude of change

Following Schlechtweg et al. (2020), the gold change score is defined as $f^*(w) = f(P^*, Q^*)$ computed from the gold SFDs X^* and Y^*.

For comparison, we used five baselines in this evaluation. As a baseline change score for the form-based approach, following Periti and Tahmasebi (2024), we used the average pairwise distance (APD) denoted as $f_{\mathrm{APD}}(w)$. As a baseline change score for the sense-based approach, we used $f_{\mathrm{WiDiD}}(w) = f(\hat{P}, \hat{Q})$ computed from the estimated SFDs $\hat{X}$ and $\hat{Y}$ in WiDiD, and also the APD between sense prototypes (APDP) yielded by clustering in WiDiD, denoted as $f_{\mathrm{APDP}}(w)$. Furthermore, we used the standard OT distance denoted as $f_{\mathrm{OT}}(w)$. We also defined a metric $f_{\mathrm{LDR}}(w)$ based on LDR of each usage instance. See Appendix G.1 for the details of the baseline metrics.

Table 3 shows the Spearman rank correlation coefficients between the gold change score f^* and the change score produced by each method. The metric f_{APD}, which directly utilizes embeddings, achieves the highest performance. However, the change score based on SUS demonstrates comparable performance, confirming that SUS effectively captures word-level semantic change. Furthermore, OT also achieves the highest performance, demonstrating that both OT and UOT are highly effective for detecting word-level semantic change.

Method	Approach	Spearman
g_{SUS}	SUS-based	0.55
g_{vMF}	form-based	**0.62**
g_{LDR}	form-based	0.36
g_{WiDiD}	sense-based	0.40

Table 4: Performance of methods for measuring word-level changes in semantic scope. For the complete results, refer to Table 12 in Appendix G.2.

7.2 Quantifying the change in semantic scope

Following (Giulianelli et al., 2020), the gold score for change in semantic scope is the entropy difference between P^* and Q^*, i.e., $g^*(w) = g(P^*, Q^*)$.

As a baseline score for the form-based approach, following Nagata et al. (2023), we used a metric called coverage denoted as g_{vMF}. Moreover, we defined g_{LDR} based on LDR of each usage instance. As a sense-based approach, we used $g_{\mathrm{WiDiD}} = g(\hat{P}, \hat{Q})$ computed from the estimated SFDs in WiDiD. See Appendix G.2 for the details of the baseline metrics.

Table 4 shows the Spearman rank correlation coefficients between the gold change score g^* and the change score calculated by each method. The form-based approach achieves the highest performance, effectively capturing the broadening or narrowing of meaning. The change score based on SUS outperforms the sense-based method, demonstrating a certain level of validity.

8 Conclusion

We applied UOT to sets of contextualized embeddings of a target word in a diachronic corpus pair. By leveraging the excess and deficit in the alignment between usage instances, we proposed a new measure called SUS for each usage instance to quantify changes in the frequency of word usage associated with its sense. The effectiveness of SUS was evaluated through experiments on semantic change detection tasks.

Limitations

- While SUS can detect detailed changes in individual usage instances, form-based methods

that directly utilize contextualized word embeddings slightly outperform the SUS-based method for quantifying word-level semantic change.

- The SUS value depends on the hyperparameter λ used in UOT. Since the optimal λ value may vary depending on the data, further investigations are needed to establish effective λ tuning strategies.

- In this paper, we used only the DWUG dataset for English. Experiments on other languages remain a topic for future work.

- Since UOT is applied under the constraint of handling transport between two probability distributions, changes in the total occurrence counts of a target word across a diachronic corpus pair are not considered. In other words, UOT focuses on the relative frequency of word usage associated with a sense within each corpus. Additionally, the dataset used in this study does not provide effective information about the total occurrence counts of the target words. These limitations could potentially be addressed in the future through extensions of UOT and the acquisition of datasets that include information on total occurrence counts.

Ethics Statement

This study complies with the ACL Ethics Policy.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments and suggestions. This study was partially supported by JSPS KAKENHI 22H05106, 23H03355, 23K24910, JST FOREST JPMJFR2331, and JST CREST JPMJCR21N3.

References

Taichi Aida and Danushka Bollegala. 2023. Unsupervised semantic variation prediction using the distribution of sibling embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6868–6882, Toronto, Canada. Association for Computational Linguistics.

Taichi Aida, Mamoru Komachi, Toshinobu Ogiso, Hiroya Takamura, and Daichi Mochihashi. 2021. A comprehensive analysis of PMI-based models for measuring semantic differences. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 21–31, Shanghai, China. Association for Computational Linguistics.

Yuki Arase, Han Bao, and Sho Yokoi. 2023. Unbalanced optimal transport for unbalanced word alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3966–3986. Association for Computational Linguistics.

Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. 2005. Clustering on the unit hypersphere using von mises-fisher distributions. *J. Mach. Learn. Res.*, 6:1345–1382.

Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.

Laetitia Chapel, Rémi Flamary, Haoran Wu, Cédric Févotte, and Gilles Gasso. 2021. Unbalanced optimal transport through non-negative penalized linear regression. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 23270–23282.

Yimeng Chen, Yanyan Lan, Ruibin Xiong, Liang Pang, Zhiming Ma, and Xueqi Cheng. 2020. Evaluating natural language generation via unbalanced optimal transport. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3730–3736. ijcai.org.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al. 2021. Pot: Python

optimal transport. *Journal of Machine Learning Research*, 22(78):1–8.

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, ACL 2020, Online, July 5-10, 2020*, pages 3960–3973. Association for Computational Linguistics.

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

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*, pages 625–635. ACM.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 957–966. JMLR.org.

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, SemEval@COLING 2020, Barcelona (online), December 12-13, 2020*, pages 126–134. International Committee 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, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4642–4652. Association for Computational Linguistics.

Ryo Nagata, Hiroya Takamura, Naoki Otani, and Yoshifumi Kawasaki. 2023. Variance matters: Detecting semantic differences without corpus/word alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15609–15622, Singapore. Association for Computational Linguistics.

Francesco Periti, Alfio Ferrara, Stefano Montanelli, and Martin Ruskov. 2022. What is done is done: an incremental approach to semantic shift detection. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, LChange@ACL 2022, Dublin, Ireland, May 26-27, 2022*, pages 33–43. Association for Computational Linguistics.

Francesco Periti and Stefano Montanelli. 2024. Lexical semantic change through large language models: a survey. *ACM Comput. Surv.*, 56(11):282:1–282:38.

Francesco Periti and Nina Tahmasebi. 2024. A systematic comparison of contextualized word embeddings for lexical semantic change. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4262–4282, Mexico City, Mexico. Association for Computational Linguistics.

Mohammad Taher Pilehvar and José Camacho-Collados. 2019. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1267–1273. Association for Computational Linguistics.

Marko Pranjic, Kaja Dobrovoljc, Senja Pollak, and Matej Martinc. 2024. Semantic change detection for slovene language: a novel dataset and an approach based on optimal transport. *CoRR*, abs/2402.16596.

David Rother, Thomas N. Haider, and Steffen Eger. 2020. CMCE at semeval-2020 task 1: Clustering on manifolds of contextualized embeddings to detect historical meaning shifts. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING 2020, Barcelona (online), December 12-13, 2020*, pages 187–193. International Committee for Computational Linguistics.

Dominik Schlechtweg, Pierluigi Cassotti, Bill Noble, David Alfter, Sabine Schulte im Walde, and Nina Tahmasebi. 2024. More dwugs: Extending and evaluating word usage graph datasets in multiple languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 16-16, 2024*, pages 14379–14393. Association for Computational Linguistics.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic usage relatedness (durel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 169–174. Association for Computational Linguistics.

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.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. 2012. *Density ratio estimation in machine learning*. Cambridge University Press.

Kyle Swanson, Lili Yu, and Tao Lei. 2020. Rationalizing text matching: Learning sparse alignments via optimal transport. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5609–5626. Association for Computational Linguistics.

Zihao Wang, Datong Zhou, Ming Yang, Yong Zhang, Chenglong Rao, and Hao Wu. 2020. Robust document distance with wasserstein-fisher-rao metric. In *Proceedings of The 12th Asian Conference on Machine Learning, ACML 2020, 18-20 November 2020, Bangkok, Thailand*, volume 129 of *Proceedings of Machine Learning Research*, pages 721–736. PMLR.

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, pages 673–681. ACM.

Sho Yokoi, Ryo Takahashi, Reina Akama, Jun Suzuki, and Kentaro Inui. 2020. Word rotator's distance. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2944–2960. Association for Computational Linguistics.

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, LChange@ACL 2022, Dublin, Ireland, May 26-27, 2022*, pages 149–164. Association for Computational Linguistics.

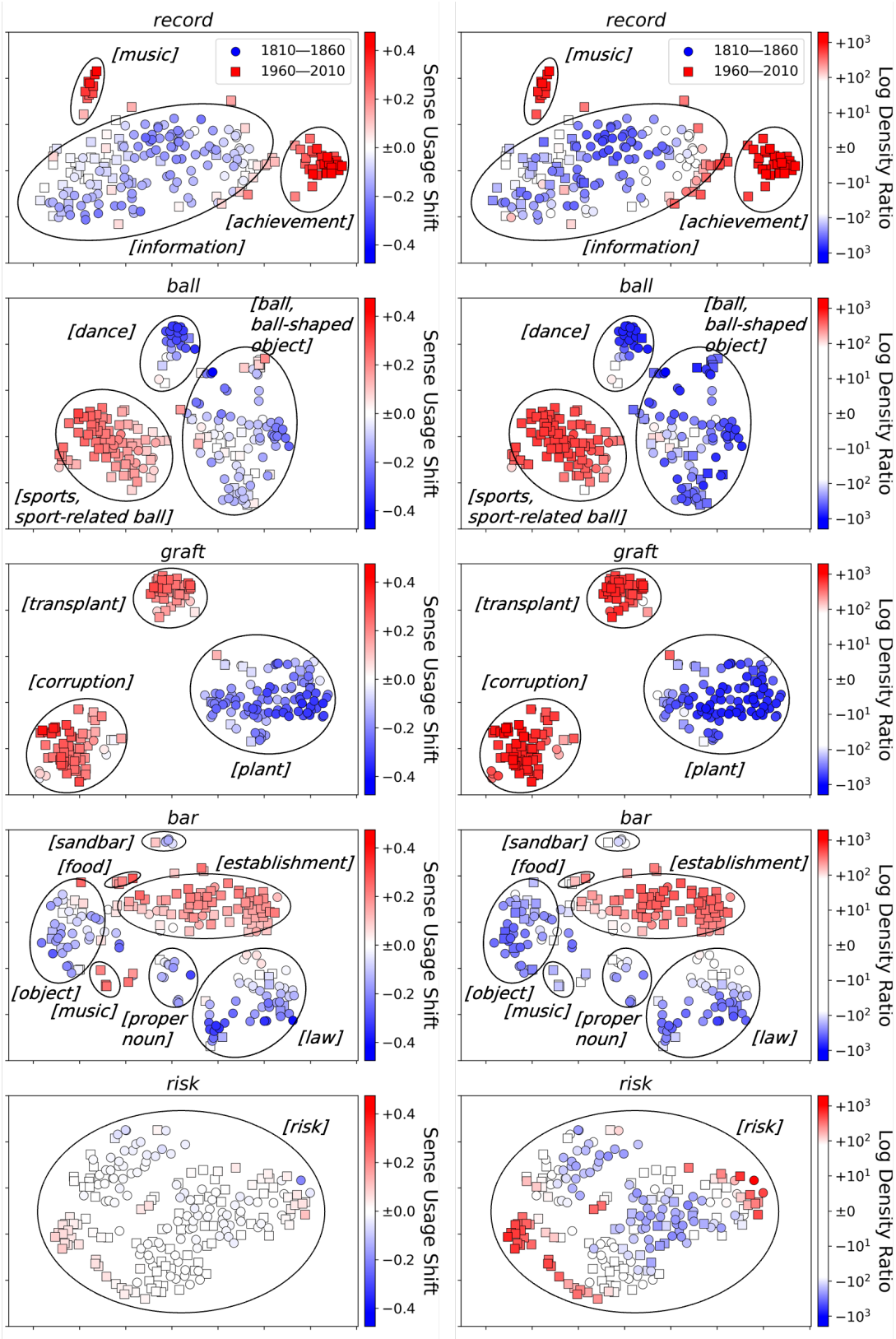
Xu Zhao, Zihao Wang, Yong Zhang, and Hao Wu. 2020. A relaxed matching procedure for unsupervised BLI. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3036–3041. Association for Computational Linguistics.

A Further Illustrative Examples

Visualizing semantic change with SUS and LDR. We selected five target words: *record*, *ball*, *graft*, *bar*, and *risk*. The semantic change visualizations for the five target words are presented in Fig. 7. Note that the target word *risk* does not exhibit semantic change in the dataset. For color-coding each instance based on its values (SUS or LDR), the maximum of the 95% percentiles and the minimum of the 30% percentiles of the absolute values for each target word were used as thresholds. Instances with values exceeding the upper threshold are highlighted with the darkest colors (red or blue), while those below the lower threshold are rendered in white. Additionally, since LDR takes very large values in high-dimensional settings, the transformation $\mathrm{sgn}(\mathrm{LDR})\log(1+|\mathrm{LDR}|)$ was applied before assigning colors.

Comparison of SUS and LDR Visualizations. As described in Section 4.4, Fig. 7 indicates that both SUS and LDR generally capture the sense clusters reasonably well. However, a more detailed examination highlights distinctions between the two approaches. For the target word *record*, when compared against the gold senses from the DWUG dataset shown in Fig. 5b, SUS appears to align more closely with the actual clusters than LDR. Specifically, in the rightmost portion of the cluster representing the sense *[information]*, there is a region where circles and squares overlap. LDR distinguishes this area as separate clusters more clearly than SUS does. For the target word *bar*, the sense *[music]* newly emerged in the modern corpus. SUS correctly identifies this increase in usage frequency, whereas LDR incorrectly identifies it as a frequency decrease. However, such detailed observations have not been sufficiently discussed, and caution is needed when drawing conclusions from these differences.

Another noteworthy difference lies in the interpretability of the SUS and LDR values. The SUS values range from -1 to $+1$, making it easier to interpret their magnitude. In contrast, LDR values fluctuate over a much larger range, often varying substantially, making it difficult to interpret their magnitude. For example, for the target word *risk*, which has only one sense *[risk]*, SUS correctly identifies that no semantic change has occurred. In contrast, LDR identifies frequency changes in unnecessarily small regions, dividing what should be a single cluster.



(a) SUS (b) LDR

Figure 7: t-SNE visualization of the contextualized embeddings for each usage instance of a target word in the diachronic corpus pair. The color of each instance represents its (a) SUS or (b) LDR.

	Corpus	Instance	Sense	SUS
Top 5	1960–2010	So did Sire ***Records***...	*[music]*	0.47
	1960–2010	... a team with the third-worst ***record***...	*[achievement]*	0.45
	1960–2010	... the AMCU single-season ***record***...	*[achievement]*	0.45
	1960–2010	... the indoor ***record*** of 24-2 1/2...	*[achievement]*	0.44
	1960–2010	... the late Steve Prefontaine's American ***record***...	*[achievement]*	0.44
Bottom 5	1810–1860	This ***record*** shall be read at the commencement...	*[information]*	-0.22
	1810–1860	... in the comprehensive ***records*** of philosophy...	*[information]*	-0.23
	1810–1860	... interpretations of the Mosaic ***record***...	*[information]*	-0.24
	1810–1860	... the ***records*** of a professed revelation...	*[information]*	-0.24
	1810–1860	... the ***record*** of whose wisdom is included in...	*[information]*	-0.25

(a) *record*

	Corpus	Instance	Sense	SUS
Top 5	1960–2010	... by teaching Wagner a palm ***ball***.	*[sport-related ball]*	0.31
	1960–2010	... flip the ***ball*** to a trailing halfback...	*[sport-related ball]*	0.27
	1960–2010	... a safety pass, and if he gets the ***ball***...	*[sport-related ball]*	0.27
	1960–2010	... pass the ***ball*** to a spot...	*[sport-related ball]*	0.26
	1960–2010	... you run that ***ball*** again you're out...	*[sport-related ball]*	0.26
Bottom 5	1810–1860	... at a tea-party, or a ***ball***...	*[dance]*	-0.32
	1810–1860	... at, a ***ball***, or dance...	*[dance]*	-0.32
	1810–1860	I now began to attend ***balls***...	*[dance]*	-0.34
	1810–1860	It is a masked ***ball***...	*[dance]*	-0.34
	1810–1860	... keep up the ***ball*** of conversation...	*[ball, ball-shaped]*	-0.42

(b) *ball*

	Corpus	Instance	Sense	SUS
Top 5	1960–2010	... political wheeling and dealing and ***graft***...	*[corruption]*	0.40
	1960–2010	... sweetheart contracts, and outright ***graft***.	*[corruption]*	0.38
	1960–2010	... there is less ***graft*** in the police department...	*[corruption]*	0.36
	1960–2010	... an average person would consider to be ***graft***.	*[corruption]*	0.34
	1960–2010	... political ***graft*** is rampant in every department...	*[corruption]*	0.34
Bottom 5	1810–1860	... between the stock and ***graft***... from a lofty tree...	*[plant]*	-0.33
	1810–1860	In the following spring, the ***grafted*** trees...	*[plant]*	-0.33
	1810–1860	... ***grafts*** of a few varieties inserted at standard height...	*[plant]*	-0.33
	1810–1860	... which have been ***grafted*** late in spring...	*[plant]*	-0.35
	1810–1860	... a tree but one year from the ***graft***...	*[plant]*	-0.35

(c) *graft*

	Corpus	Instance	Sense	SUS
Top 5	1960–2010	... from the snack ***bar***...	*[establishment]*	0.29
	1960–2010	He'd lined up the bottles on the ***bar***...	*[establishment]*	0.26
	1960–2010	... but since the window lacks scroll ***bars***...	*[object]*	0.26
	1960–2010	... the oboe to rest for a certain number of ***bars***...	*[music]*	0.24
	1960–2010	Champagne, hors d'oeuvres, a five-course dinner, open ***bar***...	*[establishment]*	0.24
Bottom 5	1810–1860	... gentlemen of the ***bar***... constant attendance upon courts...	*[law]*	-0.33
	1810–1860	... some member of the ***bar***...	*[law]*	-0.34
	1810–1860	Before a legal tribunal... at the ***bar*** of public opinion...	*[law]*	-0.35
	1810–1860	... before the ***bar*** of Judgment...	*[law]*	-0.35
	1810–1860	***bar*** of the House, against the matter of the charges...	*[law]*	-0.40

(d) *bar*

	Corpus	Instance	Sense	SUS
Top 5	1960–2010	... to take ***risks*** with executives...	*[risk]*	0.11
	1960–2010	Being overweight increases your ***risk***...	*[risk]*	0.090
	1960–2010	... people at high ***risk*** of breast cancer...	*[risk]*	0.090
	1960–2010	... an increased ***risk*** of type 2 diabetes.	*[risk]*	0.064
	1960–2010	... have an increased ***risk*** of breast cancer.	*[risk]*	0.060
Bottom 5	1810–1860	... at the ***risk*** of holding him too long from...	*[risk]*	-0.058
	1810–1860	... at the ***risk*** of her life...	*[risk]*	-0.062
	1810–1860	At the ***risk*** of repeating what may be already quite familiar...	*[risk]*	-0.064
	1810–1860	... at the ***risk*** of arguing ourselves unknown...	*[risk]*	-0.069
	1810–1860	... must charge interest and ***risk***...	*[risk]*	-0.20

(e) *risk*

Table 5: The top 5 and bottom 5 usage instances of the target words based on SUS values.

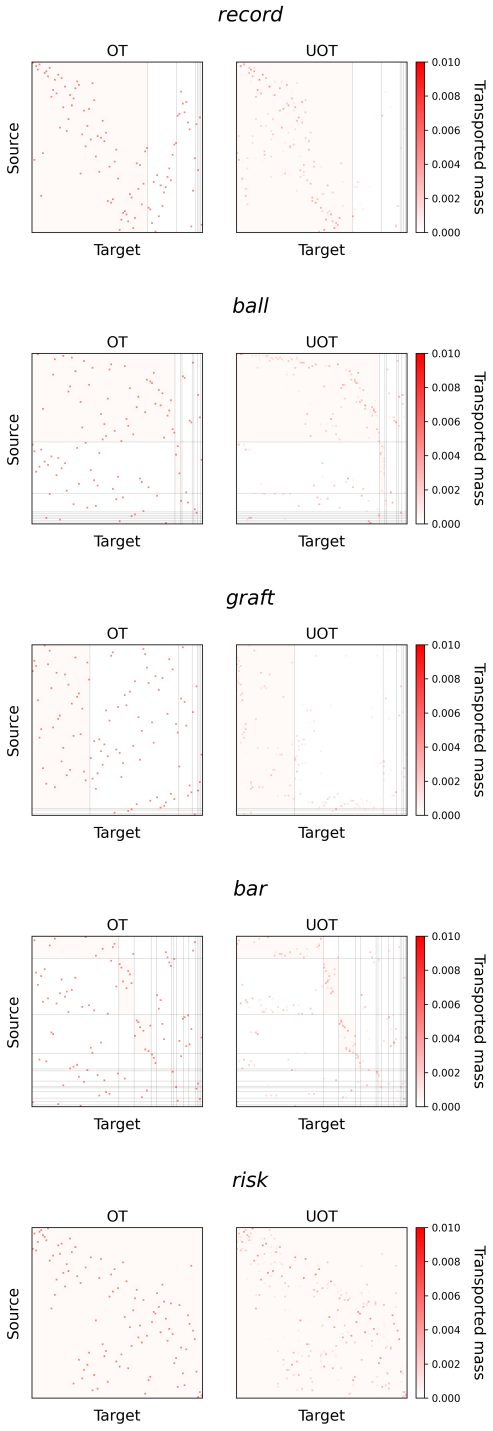


Figure 8: The transportation matrices of OT and UOT for contextualized embeddings from old and modern corpora. Red-shaded areas indicate alignments between instances in the same word sense across the two corpora, while white areas indicate alignments between different senses. By focusing on the white blocks, it is observed that OT conducts transportation across different senses, whereas UOT reduces such transportation by allowing alignment discrepancy.

Instances with high or low SUS values. Additionally, Table 5 shows the usage instances of the target words with the top 5 and bottom 5 SUS values. For example, Table 5a presents the instances of *record* with high and low SUS values. The senses of the top 5 instances correspond to those whose usage frequencies have increased in the modern corpus, such as *[music]* and *[achievement]*. Conversely, Table 5b presents the instances of *ball*, where the senses of the bottom 5 instances correspond to those whose usage frequencies have decreased in the modern corpus, including *[dance]* and *[ball, ball-shaped object]*. As discussed in Section 4.3, the meaning of *record* has broadened while the meaning of *ball* has narrowed across the two corpora. Note that word usage in the word sense *[dance]* remains common today. However, our analysis focuses on comparing the given two corpora which have the same number of instances sampled from a larger corpus. Therefore, as mentioned in the Limitations section, while we can quantify changes in the relative frequency of word usage for specific senses, we cannot capture changes in the total count of word occurrences.

Difference between OT and UOT. Figure 8 visualizes the transportation matrices of OT and UOT between the sets of the contextualized embeddings from the old and modern corpora for several target words. The usage instances are sorted according to the gold senses provided in DWUG. Within each sense, the instances are further ordered by their x-values in their t-SNE visualizations. Lines indicate the boundaries between different senses. The (i, j) element belongs to the red-shaded blocks if the senses of s_i and t_j are the same, and it belongs to the white blocks if the senses are different. In the case of OT in Fig. 8, it is observed that a substantial amount of transport occurs within the white blocks. This indicates that alignments are conducted between the instances with different senses across the two corpora. As described in Section 3.3, this happens because OT enforces a balanced alignment between the instances from the two corpora, which fails to fully capture semantic change at the instance level. On the other hand, for UOT, the transported mass within the white blocks is observed to be small, indicating that transportation across instances with different senses is avoided. This is because, as discussed in Section 4.1, UOT allows for excess or deficit of alignment. Thus, UOT more effectively captures semantic change at

the instance level.

B Details of Log-Density Ratio

Computation. In our experiments, we directly computed LDR from Eq. (8) of Nagata et al. (2023). Specifically, we calculated the approximate maximum likelihood estimator (MLE) of the concentration parameter κ for the von Mises-Fisher (vMF) distribution for the old and modern corpora as follows: Let $\{\boldsymbol{u}_i\}_{i=1}^m \subset \mathbb{R}^d$ denote the set of embeddings in the old corpus and define $\tilde{\boldsymbol{u}}_i = \boldsymbol{u}_i/\|\boldsymbol{u}_i\|$. Assume that $\{\tilde{\boldsymbol{u}}_i\}_{i=1}^m$ follow a vMF distribution with mean direction parameter $\boldsymbol{\mu}_S$ and concentration parameter κ_S. Let $\ell = \left\|\frac{1}{m}\sum_i \tilde{\boldsymbol{u}}_i\right\|$. Then, the maximum likelihood estimator (MLE) of $\boldsymbol{\mu}_S$ is given by $\frac{1}{\ell m}\sum_i \tilde{\boldsymbol{u}}_i$. While the MLE of κ_S does not have a closed-form solution, Banerjee et al. (2005) provide the following approximation: $\kappa_S \approx \frac{\ell(d-\ell^2)}{1-\ell^2}$. After calculating the MLEs, we derived the corresponding vMF density function using `vonmises_fisher` in `scipy`. We applied the same procedure to the modern embeddings $\{\boldsymbol{v}_j\}_{j=1}^n$ by computing their normalized forms $\tilde{\boldsymbol{v}}_j = \boldsymbol{v}_j/\|\boldsymbol{v}_j\|$ and estimating the parameters $\boldsymbol{\mu}_T$ and κ_T of the associated vMF distribution.

Representativeness. In Nagata et al. (2023), an approximation of LDR with the constant term omitted is introduced as *representativeness* in Eq. (3), which is proposed as a measure for extracting typical word instances. Due to the omission of the constant term, the sign of *representativeness* does not indicate whether usage has increased or decreased.

C Detailed Settings for UOT

C.1 Computation of UOT

The Unbalanced Optimal Transport (UOT) problem is formulated as (2). In general, if the objective function is convex, the Majorization-Minimization (MM) algorithm can be applied. Specifically, when D_1 and D_2 are Bregman divergences[7], they are convex with respect to their first arguments, making the objective function convex. This implies that the optimal solution to (2) can be computed using the MM algorithm (Chapel et al., 2021). The standard OT is recovered when $\lambda_1 = \lambda_2 \to \infty$.

C.2 Determining the range for hyperparameter tuning

The hyperparameter $\lambda_1 = \lambda_2 = \lambda$ in UOT (2) needs to be tuned on a validation set during the performance evaluations conducted in Sections 6 and 7. To determine the range of λ values to explore, we initially evaluated the performance using all target words without splitting the validation and test sets. The results of varying λ across the values 1, 10, 100, and 1000 are presented in Table 6. Note that additional metrics f_1, f_2, f_3 and g_1 for quantifying the degree of semantic change have been proposed as defined in Appendix G.

Additionally, since the range of SUS values for each target word varies depending on λ, the threshold θ used in f_2 and g_1 is determined as a proportion of the maximum absolute SUS value within a given set of target words (valid, test, or all). Specifically, let M denote the maximum value of $|\mathrm{SUS}(s_i)|$ or $|\mathrm{SUS}(t_j)|$ for $i = 1, \ldots, m$, $j = 1, \ldots, n$, and $w \in W$, where W is the set of target words. The threshold is then defined as $\theta = Mr$, where r is the proportion to explore. In other words, instead of directly tuning θ, we search for r. Initially, r is varied across values $0.1, 0.2, \ldots, 0.9$ to establish a baseline range. Table 6 presents the optimal r for each value of λ.

Based on these results, when the only hyperparameter to be tuned was λ, we explored $\lambda \in \{10, 20, 50, 100, 200, 500, 1000\}$. When both λ and r were tuned, we varied $\lambda \in \{10, 100, 1000\}$ and $r \in \{0.4, 0.6, 0.8\}$.

C.3 Details of hyperparameter tuning in Sections 6 and 7

The hyperparameter λ is tuned on the validation set for each performance evaluation task. The detailed procedure is outlined as follows:

Tuning of λ. The 46 target words in the DWUG dataset are randomly split into validation and test sets in an 8:2 ratio. For the validation set, the optimal λ is determined by exploring $\lambda \in \{10, 20, 50, 100, 200, 500, 1000\}$. In the case of WiDiD, a hyperparameter called damping (ranging in $[0.5, 1.0)$) is tuned by exploring $\{0.5, 0.6, 0.7, 0.8, 0.9\}$. Subsequently, performance evaluation is conducted on the test set. This validation-test splitting is repeated 100 times, and the final performance of each method is reported as the average test performance across all splits.

[7] Examples include KL divergence, L1 error, and L2 error.

	Gold metric	Proposed metric	λ			
			1	10	100	1000
instance-level	τ^*	τ_{SUS}	0.31	0.42	**0.50**	0.49
word-level	f^*	f_{SUS}	0.54	0.73	**0.76**	**0.76**
		f_1	0.42	**0.73**	0.72	0.70
		f_2	0.48 (0.9)	**0.78** (0.8)	0.75 (0.2)	0.74 (0.2)
		f_3	-0.55	-0.23	**0.76**	**0.76**
	g^*	g_{SUS}	-0.06	0.57	**0.59**	0.56
		g_1	0.00 (0.9)	0.54 (0.8)	**0.57** (0.5)	0.55 (0.4)

Table 6: Performance of each task using all target words. The performance in the tasks using gold metrics f^*, g^*, and τ^* is reported as Spearman rank correlations. For methods involving the threshold θ, the optimal value of r is provided in parentheses alongside the performance.

Proposed metric	Selected λ or (λ, r)	Times
τ_{SUS}	100	73
f_{SUS}	20	51
f_1	20	90
f_2	(10, 0.8)	78
f_3	100	46
g_{SUS}	50	54
g_1	(10, 0.8)	53

Table 7: The most frequently selected hyperparameters for the proposed instance-level and word-level metrics in the validation set, where train-validation splits are conducted 100 times.

Tuning of λ and r. The 46 target words in the DWUG dataset are randomly split into validation and test sets in a ratio of 8:2. For the validation set, the optimal λ and r are determined by exploring $\lambda \in \{10, 100, 1000\}$ and $r \in \{0.4, 0.6, 0.8\}$. The subsequent procedure follows the same steps as described above.

Selected hyperparameters. Table 7 shows the most frequently selected hyperparameter values for each evaluation task.

D Details of DWUG

In this appendix, we provide additional details on how DWUG (Schlechtweg et al., 2021, 2024) constructs the gold senses for a target word, which were briefly introduced in Section 5. In the DWUG dataset[8], a sufficient number of usage instance pairs for each target word are annotated by human annotators with a four-level similarity score, called the DURel relatedness scale (Schlechtweg et al., 2018). DWUG clusters the instances, represented as vertices in a network, based on the similarity scores, which are used as edge weights. Through this clustering, the sense of each usage instance is identified. In other words, DWUG can be interpreted as one of a sense-based approach via human-annotated similarities, not using the similarities obtained by contextualized embeddings. Since the clustering algorithm is applied, the gold SFDs may contain noise senses as shown in Table 8.

Due to the specification of the DWUG dataset, a small number of usage instances for each target word are not assigned a sense. In this paper, we refer to such senses as *undefined*. However, in cases where the sense can be easily identified by the authors of this paper, the missing senses are annotated, such as *[music]* in Table 1.

E Details of XL-LEXEME

We describe how XL-LEXEME[9] (Cassotti et al., 2023) calculates the embeddings of target words. XL-LEXEME is a model obtained by fine-tuning XLM-RoBERTa (Conneau et al., 2020) on Word-in-Context (Pilehvar and Camacho-Collados, 2019) and it has 561M parameters. In the DWUG dataset, each usage instance of a target word includes positional information specifying the position of the target word within its context. XL-LEXEME takes this context and positional information as input to compute the embedding for the target word. Specifically, special tokens are inserted before and after the target word in the context, and the embedding of the target word is computed as the mean of the embeddings of all tokens in the context.

F Details of Experiment in Section 6

For computing SUS, the hyperparameter λ in (2) was tuned by splitting the 46 target words in the

[8]The dataset is licensed under the CC BY-ND 4.0. https://zenodo.org/records/14028531

[9]https://huggingface.co/pierluigic/xl-lexeme

Corpus	Instance	Sense	Gold SFD
1810–1860	... in all the ***records*** of sorrow...	*[information]*	$X^* = (99, 0, 0, 0, 0, 0, 0)$
	... in the ***record*** of any age or country...	*[information]*	
	... the historic ***records*** of Christianity...	*[information]*	
1960–2010	... this is the ***record*** of my life...	*[information]*	$Y^* = (64, 17, 11, 1, 1, 1, 1)$
	... single-season ***record*** held by...	*[achievement]*	
	The ***record*** labels' new service...	*[music]*	

Table 8: Three usage instances each from the old and modern corpora for the target word *record* included in the DWUG dataset. In DWUG, the gold Sense Frequency Distribution (SFD) is given. It represents the frequency of the target word usage with each sense. Upon reviewing the senses corresponding to each index, we identified them as *[information]*, *[achievement]*, *[music]*, and four noise senses.

Method	Approach	Stable (25)	Changed (21)	Overall (46)
τ_{SUS}	SUS-based	**0.21**	**0.61**	**0.46**
τ_{LDR}	form-based	0.10	0.34	0.40
τ_{WiDiD}	sense-based	0.15	0.41	0.31

Table 9: Performance of methods for predicting the instance-level change score. Overall corresponds to 'Instance-level' in Table 2.

Method	Approach	Stable (25)	Changed (21)	Overall (46)
τ_{SUS}	SUS-based	**0.70**	0.85	0.83
τ_{LDR}	form-based	0.64	0.38	0.70
τ_{WiDiD}	sense-based	0.54	**0.93**	**0.84**

Table 10: Performance of methods for predicting the sense-level change score. Overall corresponds to 'Sense-level' in Table 2.

DWUG dataset into a validation set and a test set, using the validation set for optimization. For further details, refer to Appendix C. The selected values of λ are provided in Table 7 therein.

Gold change score. The gold graded change score for a usage instance of a target word is defined as $\tau^*(\cdot) = \tau(\cdot; X^*, Y^*)$, where X^* and Y^* are the gold SFDs from DWUG.

Proposed change score. We used the SUS value for each instance as an estimator of the gold graded change score τ^*. Specifically, we defined $\tau_{\text{SUS}}(\cdot) = \text{SUS}(\cdot)$.

Baseline change score. Using the SFDs $\hat{X}$ and $\hat{Y}$ estimated by the sense-based approach WiDiD, we calculated an estimate of τ^* as $\tau_{\text{WiDiD}}(\cdot) = \tau(\cdot; \hat{X}, \hat{Y})$. Moreover, we used LDR for each usage instance directly, denoted as $\tau_{\text{LDR}}(\cdot) = \text{LDR}(\cdot)$.

Results. In the DWUG dataset, each target word is annotated with a label indicating whether its meaning has changed, based on old and modern SFDs. We divide the test set words according to this change annotation and report the instance-level and sense-level performances for each group in Table 9 and Table 10, respectively. The overall score is computed without splitting the test set and corresponds to Table 2. Hyperparameter tuning was performed using the validation set without splitting words into stable or changed categories. Performance evaluation was then conducted separately for stable and changed words in the test set.

For stable words, instance-level evaluation performs quite poorly across all methods. In stable words, semantic change is minimal, but the values of each metric still fluctuate, and even small variations in these values can result in significant changes in rankings. Therefore, instance-level rank correlation is not the most appropriate evaluation metric in this context. In any way, SUS demonstrates superior instance- and sense-level performance compared to LDR not only overall but also within both the stable and changed word categories.

Method	Approach	Spearman
f_{SUS}	SUS-based	0.69
f_1	SUS-based	0.68
f_2	SUS-based	0.68
f_3	SUS-based	0.69
f_{OT}	form-based	**0.71**
f_{APD}	form-based	**0.71**
f_{LDR}	form-based	0.31
f_{WiDiD}	sense-based	0.45
f_{APDP}	sense-based	0.51

Table 11: Performance of methods for measuring the magnitude of word-level semantic change.

G Details of Experiment in Section 7

G.1 Quantifying the magnitude of word-level semantic change

Gold change score. Following Schlechtweg et al. (2020), the gold change score is calculated by the metric f defined as $f^*(w) = f(P^*, Q^*)$.

Proposed change scores. Although in Section 4.5, we only presented f_{SUS} for brevity, we also designed other metrics:

$$f_{\mathrm{SUS}}(w) = \left| \frac{1}{m}\sum_{i=1}^{m}\alpha_i - \frac{1}{n}\sum_{j=1}^{n}\beta_j \right|,$$
$$f_1(w) = \sum_i |\alpha_i| + \sum_j |\beta_j|,$$
$$f_2(w;\theta) = -\sum_{i:\alpha_i<-\theta}\alpha_i + \sum_{j:\beta_j>\theta}\beta_j,$$
$$f_3(w) = \sum_{i,j} C_{ij}T_{ij}.$$

Here, we defined $\alpha_i = \mathrm{SUS}(s_i)$ and $\beta_j = \mathrm{SUS}(t_j)$. The second measure, f_1, quantifies the magnitude of semantic change by summing the micro-level changes across all instances, as represented by the SUS values. The third measure, $f_2(\cdot;\theta)$, focuses on instances with large SUS values, emphasizing significant changes by introducing a threshold θ to filter out smaller SUS values. The fourth measure, f_3, corresponds to the transportation distance, which is the total cost of aligning the embeddings. Note that f_{APD} is derived from f_3 by setting $T_{ij} = 1/mn$. The threshold θ is a hyperparameter. For details on the tuning process for λ and θ, see Appendix C.

Baseline change scores. As a form-based approach, following Periti and Tahmasebi (2024), we used the average pairwise distance (APD) based on cosine distance between the embeddings $\{\boldsymbol{u}_i\}_i$ and $\{\boldsymbol{v}_j\}_j$[10]:

$$f_{\mathrm{APD}}(w) = \frac{1}{mn}\sum_{i,j}(1-\cos(\boldsymbol{u}_i,\boldsymbol{v}_j)).$$

Moreover, we used standard OT distance, which was defined as

$$f_{\mathrm{OT}}(w) = \sum_{i,j} C_{ij}T_{ij},$$

where $\boldsymbol{T}$ was the optimal solution for the standard OT problem (1). We also defined a metric based on LDR for each usage instance. Let $\mathrm{LDR}(s_i)$ denote the value of LDR for s_i. Following the definition of f_{SUS}, we defined f_{LDR} as

$$f_{\mathrm{LDR}}(w) = \left| \frac{1}{m}\sum_{i=1}^{m}\mathrm{LDR}(s_i) - \frac{1}{n}\sum_{j=1}^{n}\mathrm{LDR}(t_j) \right|.$$

As a sense-based approach, we used WiDiD to estimate the normalized SFDs $\hat{P}$ and $\hat{Q}$ for the old and modern corpora, respectively. Using these estimated SFDs, we defined the two metrics f_{WiDiD} and f_{APDP}. The former was defined by $f_{\mathrm{WiDiD}}(w) = \mathrm{JSD}(\hat{P},\hat{Q})$. The latter was defined as the APD between sense prototypes (APDP). APDP is calculated as follows: For each sense k, compute the mean embeddings $\boldsymbol{\mu}_k$ and $\boldsymbol{\nu}_k$ for the old and modern corpora, respectively. The APDP measures the APD between the sets of sense prototypes $\{\boldsymbol{\mu}_k\}_k$ and $\{\boldsymbol{\nu}_k\}_k$. Following Periti and Tahmasebi (2024), we used the Canberra distance as the metric for APD.

Method	Approach	Spearman
g_{SUS}	SUS-based	0.55
g_1	SUS-based	0.45
g_{vMF}	form-based	**0.62**
g_{LDR}	form-based	0.36
g_{WiDiD}	sense-based	0.40

Table 12: Performance of methods for measuring word-level changes in semantic scope.

Results. Table 11 shows the Spearman rank correlation between the gold change score f^* and the change score produced by each method. The form-based approach that directly uses embeddings achieves the highest performance. However, the change scores based on SUS demonstrate comparable performance, confirming that SUS effectively captures word-level semantic change.

G.2 Quantifying the word-level change in semantic scope

Gold change score. Following Giulianelli et al. (2020), the gold change score is defined as the entropy difference between P^* and Q^*, i.e., $g^*(w) = g(P^*,Q^*)$.

Proposed change scores. For a target word, let the variances of SUS in the old and modern corpora be $V_S = \mathrm{Var}(\{\mathrm{SUS}(s_i)\}_{i=1}^m)$ and $V_T = \mathrm{Var}(\{\mathrm{SUS}(t_j)\}_{j=1}^n)$, respectively. In Section 4.5, we only showed g_{SUS}, whereas another metric was

[10]Here, if we normalize the embeddings as $\tilde{\boldsymbol{u}}_i = \boldsymbol{u}_i/\|\boldsymbol{u}_i\|$, $\tilde{\boldsymbol{v}}_j = \boldsymbol{v}_j/\|\boldsymbol{v}_j\|$, and compute the mean vectors $\tilde{\boldsymbol{\mu}} = \sum_{i=1}^m \tilde{\boldsymbol{u}}_i/m$ and $\tilde{\boldsymbol{\nu}} = \sum_{j=1}^n \tilde{\boldsymbol{v}}_j/n$, the APD can be interpreted as a kind of distance between these mean vectors:$f_{\mathrm{APD}}(w) = 1 - \tilde{\boldsymbol{\mu}}^\top\tilde{\boldsymbol{\nu}}$.

also designed:

$$g_{\mathrm{SUS}}(w) = \log \frac{V_T}{V_S},$$
$$g_1(w;\theta) = \sum_{i:\alpha_i<-\theta} \alpha_i + \sum_{j:\beta_j>\theta} \beta_j.$$

Here, $\alpha_i = \mathrm{SUS}(s_i)$ and $\beta_j = \mathrm{SUS}(t_j)$. The SUS value for each instance of a target word reflects the extent to which the usage in its sense has increased or decreased. Based on this, we aimed to quantify the change in semantic scope by summing the SUS values across all instances of the word. However, directly summing all SUS values results in a value of 0, i.e., $\sum_i \alpha_i + \sum_j \beta_j = 0$, due to the definition of SUS. To address this, we introduced a threshold θ to focus only on SUS values with a significant impact, calculating the sum of those values exceeding the threshold. The threshold θ is a hyperparameter. For details on the tuning process for λ and θ, refer to Appendix C.

Baseline change scores. As a form-based approach, following Nagata et al. (2023), we used a metric called coverage, defined as:

$$g_{\mathrm{vMF}}(w) = \log \frac{\kappa_S}{\kappa_T},$$

where κ_S and κ_T are known as concentration parameters, representing the spread of the distributions $\{\boldsymbol{u}_i\}_i$ and $\{\boldsymbol{v}_j\}_j$, respectively[11]. We also defined a metric based on LDR for each usage instance. Let the variances of LDR in the old and modern corpora be $U_S = \mathrm{Var}(\{\mathrm{LDR}(s_i)\}_{i=1}^m)$ and $U_T = \mathrm{Var}(\{\mathrm{LDR}(t_j)\}_{j=1}^n)$, respectively. Following the definition of g_{SUS}, we defined g_{LDR} as

$$g_{\mathrm{LDR}}(w) = \log \frac{U_T}{U_S}.$$

As a sense-based approach, we defined the metric using the entropy difference between the normalized SFDs $\hat{P}$ and $\hat{Q}$, which are estimated by WiDiD. This metric was expressed as $g_{\mathrm{WiDiD}}(w) = g(\hat{P}, \hat{Q})$.

Results. Table 12 shows the Spearman rank correlation between the gold change score g^* and the change score calculated by each method. The form-based approach achieves the highest performance, indicating that using the values corresponding to the variance of embeddings in each time period effectively captures the broadening or narrowing of meaning. The change score based on SUS outperforms the sense-based method, demonstrating a certain level of validity.

[11]These concentration parameters are derived from the norms $\|\tilde{\boldsymbol{\mu}}\|$ and $\|\tilde{\boldsymbol{\nu}}\|$, where $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\nu}}$ are the mean vectors of the normalized embeddings $\{\tilde{\boldsymbol{u}}_i\}_i$ and $\{\tilde{\boldsymbol{v}}_j\}_j$, respectively. The embeddings are assumed to follow a von Mises-Fisher (vMF) distribution, and the concentration parameters κ_S and κ_T represent the reciprocal of variance of these distributions.

Method	Approach	Spearman
f_{SUS}	SUS-based	**0.61**
f_1	SUS-based	0.60
f_2	SUS-based	0.60
f_3	SUS-based	**0.61**
f_{OT}	form-based	**0.61**
f_{APD}	form-based	**0.61**
f_{LDR}	form-based	0.36
f_{WiDiD}	sense-based	0.41
f_{APDP}	sense-based	0.52

Table 13: Performance of methods for measuring the magnitude of word-level semantic change in the Spanish DWUG dataset.

Method	Approach	Spearman
g_{SUS}	SUS-based	0.36
g_1	SUS-based	**0.45**
g_{vMF}	form-based	0.41
g_{LDR}	form-based	0.31
g_{WiDiD}	sense-based	0.33

Table 14: Performance of methods for measuring word-level changes in semantic scope in the Spanish DWUG dataset.

H Experiments on Another Dataset

Since the SUS-based metric evaluated in Section 7 may be overfitting to the English DWUG dataset (DWUG EN), we conduct an evaluation experiment for word-level semantic change detection using the Spanish DWUG dataset (Zamora-Reina et al., 2022) (DWUG ES) as an additional dataset to verify the effectiveness of SUS. The experimental setup including hyperparameter tuning follow the same procedure as in Section 7. The search scope for the hyperparameter is the same as in the experiment with DWUG EN.

Quantifying the magnitude of word-level semantic change. Table 13 shows the Spearman rank correlation between the gold change score f^* and the change score produced by each method. The SUS-based approach demonstrates performance comparable to the SOTA method f_{APD}, suggesting that SUS is also effective in measuring the degree

of semantic change in DWUG ES.

Quantifying the word-level change in semantic scope. Table 14 shows the Spearman rank correlation between the gold change score g^* and the change score calculated by each method. For g_{SUS}, when compared to the baselines, the ranking of results is consistent with that in DWUG EN, and for g_1, it outperforms g_{vMF}. This suggests that SUS is also effective in capturing changes in semantic scope in DWUG ES.