SCDTour: Embedding Axis Ordering and Merging for Interpretable Semantic Change Detection

Taichi Aida
Tokyo Metropolitan Unviersity
taichia@tmu.ac.jp

Danushka Bollegala
University of Liverpool
danushka@liverpool.ac.uk

Abstract

In semantic change detection (SCD), it is a common problem to obtain embeddings that are both interpretable and high-performing. However, improving interpretability often leads to a loss in the SCD performance, and vice versa. To address this problem, we propose SCDTour, a method that orders and merges interpretable axes to alleviate the performance degradation of SCD. SCDTour considers both (a) semantic similarity between axes in the embedding space, as well as (b) the degree to which each axis contributes to semantic change. Experimental results show that SCDTour preserves performance in semantic change detection while maintaining high interpretability. Moreover, agglomerating the sorted axes produces a more refined set of word senses, which achieves comparable or improved performance against the original full-dimensional embeddings in the SCD task. These findings demonstrate that SCDTour effectively balances interpretability and SCD performance, enabling meaningful interpretation of semantic shifts through a small number of refined axes.[1]

1 Introduction

The meanings of words shift over time due to changes in culture, society, and contexts. SCD is the task of detecting these changes automatically, which plays a vital role in aiding linguistic analysis (Kutuzov et al., 2018; Schlechtweg et al., 2020). Additionally, it also contributes to the efficient additional training of masked language models (MLMs) by identifying words whose meanings have changed over time (yu Su et al., 2022). A wide range of methods have been proposed using static (Kim et al., 2014; Hamilton et al., 2016; Dubossarsky et al., 2019; Aida et al., 2021) and contextualised word embeddings (Kutuzov and Giulianelli, 2020; Rosin et al., 2022; Rosin and Radinsky, 2022; Aida and Bollegala, 2023b,a) to enhance the performance of SCD.

Method	Dim.	Sorted	Perf.	Int.
Raw	Full (d)	✗	✓	✗
PCA	Top-k	eigenvalue	✓	✗
ICA	Top-k	skewness	✗	✓
SCDTour	Merge-k	TSP	✓	✓

Table 1: Comparison of different embedding methods in terms of dimension (**Dim.**), axis-sorting strategy (**Sorted**), SCD performance (**Perf.**), and axis interpretability (**Int.**). Merge-k represents the process of merging adjacent axes based on TSP sorting to obtain k-dimensional embeddings ($k < d$).

Recently, interpretability has emerged as a key focus in SCD. While earlier approaches prioritised improving accuracy using contextualised or static word embeddings, recent work has focused on transparency by generating definitions (Giulianelli et al., 2023; Kutuzov et al., 2024), building usage graphs (Schlechtweg et al., 2021; Ma et al., 2024), analysing embedding space structure (Nagata et al., 2023; Aida and Bollegala, 2023b), or leveraging external knowledge (Tang et al., 2023; Periti et al., 2024b; Baes et al., 2024). However, a key challenge remains: **improving interpretability often leads to reduced performance, and vice versa** (Aida and Bollegala, 2025). This trade-off limits practical applications that demand both reliable predictions and interpretable explanations.[2]

We address this issue by utilising interpretable embeddings whose axes are obtained via Independent Component Analysis (ICA). ICA has been used to derive interpretable axes in word embeddings that encode meaning-specific information (Yamagiwa et al., 2023). We propose **SCDTour**, an interpretable axis-sorting method that

[1] Source code is available at https://github.com/LivNLP/svp-tour.

[2] We define interpretability in SCD as the ability to assign human-interpretable meanings to individual embedding axes. Unlike methods that generate textual definitions or use external knowledge bases, we focus on making each dimension of the representation space interpretable, enabling direct inspection of the semantic properties captured by the model.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 14775–14785
November 4-9, 2025

Method	Categorical	Similarity	Analogy
Raw, $d = 300$	0.68	0.57	0.50
PCA			
$k = 5$	0.36	0.15	**0.02**
$k = 20$	0.49	0.23	**0.09**
$k = 100$	0.62	0.48	0.39
ICA=ICA(PCA)			
$k = 5$	0.30	0.06	0.00
$k = 20$	0.41	0.20	0.04
$k = 100$	0.60	0.46	0.35
PCA(ICA)			
$k = 5$	0.34	0.19	0.01
$k = 20$	0.42	0.39	0.04
$k = 100$	0.58	**0.53**	0.40
SCDTour (λ = 0.00) (Yamagiwa et al., 2024)			
$k = 5$	**0.40**	**0.26**	0.00
$k = 20$	**0.52**	**0.42**	0.07
$k = 100$	**0.63**	0.51	**0.46**

Table 2: The performance of GloVe embeddings. We used the pretrained `GloVe 6B` model, referred to Yamagiwa et al. (2024). **ICA(PCA)** and **PCA(ICA)** indicate that PCA/ICA is conducted for the **Raw** embeddings to obtain full-dimensional axes ($d = 300$), then ICA/PCA is performed to obtain d-dimensional embeddings.

extends prior work (Sato, 2022; Yamagiwa et al., 2024) by introducing *change-specific weights* as a novel criterion, in addition to meaning-specific weights, to investigate **whether the ICA-derived axes capture and explain semantic change of words**. SCDTour enables us to sort and merge axes into interpretable embeddings while preserving SCD performance. Experimental results show that SCDTour can obtain low-dimensional, high-performing, and interpretable representations for SCD against standard dimension reduction methods such as PCA and ICA (Table 1).

2 Method

We propose **SCDTour**, which introduces a **change-specific weight** to account for the contribution of each axis to SCD. Unlike WordTour (Sato, 2022), which sorts words based on pairwise similarity to obtain one-dimensional embeddings, and AxisTour (Yamagiwa et al., 2024), which aligns ICA axes and merges similar ones for better compression, our method incorporates change-specific signals to reorder and merge topic-like axes, thereby achieving both interpretability and SCD performance. Table 1 shows that SCDTour is the only method that achieves both axis-level interpretability and high SCD performance.

Previous work has explored how to sort axes in word embeddings to improve interpretability. WordTour (Sato, 2022) reorders n words in the d-dimensional original (Raw) static word embedding (SWE) $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n]^\mathsf{T} \in \mathbb{R}^{n \times d}$.[3] To obtain the optimal ordering σ, the task is formulated as a Travelling Salesman Problem (TSP) over words:

$$\min_{\sigma \in \mathcal{P}([n])} w(\sigma_1, \sigma_n) + \sum_{i=1}^{n-1} w(\sigma_i, \sigma_{i+1}), \quad (1)$$

where $\mathcal{P}([n])$ denotes the set of permutations of n words, and $w(i, j)$ is a weight function that quantifies semantic distance between i-th and j-th words. In WordTour, $w(i, j)$ is defined as the L1 distance between word embeddings $w(i, j) = ||\boldsymbol{x}_i - \boldsymbol{x}_j||$.

Building on this idea, AxisTour (Yamagiwa et al., 2024) proposed to sort ICA-transformed axes $\mathbf{S} = \mathbf{A}_{\mathrm{ICA}}^\mathsf{T} \mathbf{X}^\mathsf{T} \in \mathbb{R}^{d \times n}$ by their semantic similarity between the i-th and j-th axes (**meaning-specific weight**) to obtain meaning-related ordering of axes:

$$w_m(i, j) = \cos(\mathbf{v}_i, \mathbf{v}_j), \quad (2)$$

where $\mathbf{v}_i$ is the mean embeddings of the top N words in the i-th axis. In addition, AxisTour introduced a dimension reduction technique ($\mathbb{R}^{n \times d} \rightarrow \mathbb{R}^{n \times k}$) by merging adjacent axes $I_r = \{a_r, .., b_r\}, r \in \{1, ..., k\}$ along the sorted order:

$$f_r^{(\ell)} = \begin{cases} \dfrac{\gamma_\ell^\alpha}{\sqrt{\sum_{i=a_r}^{b_r} \gamma_i^{2\alpha}}} & \text{for } \ell \in I_r, \\ 0 & \text{otherwise,} \end{cases} \quad (3)$$

where γ_i denotes the skewness of the i-th axis, and $f_r^{(\ell)}$ indicates how much the ℓ-th axis contributes to the r-th reduced dimension.

To extend this approach for semantic change detection (SCD), we introduce an additional criterion: the **change-specific weight**, which evaluates how much each axis contributes to the performance on the SCD task. Formally, we define it as:

$$w_c(i, j) = ||Imp(j) - Imp(i)||, \quad (4)$$

where $Imp(i)$ is the **importance of the i-th axis** compared to all other dimensions $\{D\}$, defined as below:

$$Imp(i) = E(\mathbf{S}_{\{D\}}) - E(\mathbf{S}_{\{D\} \setminus \{i\}}). \quad (5)$$

It quantifies the drop in performance E when the i-th axis is removed. Our proposed method,

[3] WordTour focuses on constructing one-dimensional trajectories of words for interpretability. However, it is not directly applicable to SCD, which requires comparing embeddings across multiple time periods.

SCDTour, combines both criteria to produce a more informative and SCD-relevant axis ordering:

$$w(i,j) = \lambda w_c(i,j) + (1-\lambda) w_m(i,j), \quad (6)$$

where $\lambda \in [0,1]$ is a hyperparameter to control the change-specific weight. This generalises previous methods, which rely solely on semantic similarity. While interpretability is not guaranteed, our preliminary experiment shows that this merging strategy achieves comparable or better performance to standard dimension reduction methods such as PCA or ICA on word embedding benchmarks (Table 2).

3 Experiments

3.1 Settings

To evaluate the effectiveness of SCDTour, we focus on two aspects: (a) the interpretability of axes, and (b) its performance in SCD. Additionally, we investigate how the weighting parameter λ, balancing change-specific and meaning-specific weights in Equation 6, influences both aspects. Following previous work, we employ the LKH solver (Helsgaun, 2000) to solve the TSP formulated in Equation 1.

Interpretability: To assess interpretability, we use the Word Intruder Test (WIT) (Musil and Mareček, 2024). This task measures the axis coherence by introducing a semantically unrelated *intruder* word into a set of related words[4] and checks whether an evaluator can correctly identify the intruder word. Following Musil and Mareček (2024), we use Large Language Models (LLMs) to simulate human-level evaluation. Specifically, we adopt three publicly available instruction-tuned models: Llama-3.1,[5] Gemma-3,[6] and Qwen3.[7] Llama-3.1 has previously demonstrated effectiveness in a recent SCD task (Periti et al., 2024a). We also include Gemma-3 and Qwen3 to examine the robustness of our interpretability results across different LLM architectures and training strategies. We prompt the model in a zero-shot setting and post-process outputs to extract a single-word prediction. To account for randomness in generation, we report the average accuracy over five runs.

Performance: We use the standard benchmark, SemEval-2020 Task 1 (Schlechtweg et al., 2020), which provides two time-separated corpora and a list of target words.[8] Following prior works (Cassotti et al., 2023; Periti et al., 2024b; Aida and Bollegala, 2024), we mainly conduct the ranking task and measure the Spearman's correlation between semantic change scores and human ratings. In addition, we also evaluate the binary classification setting, where the goal is to decide whether a target word has changed in meaning across time periods. Similar to the WIT evaluation, we adopt three LLMs. However, the binary setting is evaluated under few-shot prompting to mitigate overly strict decisions.[9]

In our experiments, we use SWEs instead of contextualised word embeddings (CWEs), which aligns with previous studies (Sato, 2022; Yamagiwa et al., 2024; Musil and Mareček, 2024), because SWEs provide more explicit access to axis-level information. While CWEs encode sense-aware information for each token occurrence and achieve higher performance on SCD (Cassotti et al., 2023), it makes axis-level interpretation difficult due to the large number of contextualised instances. In contrast, SWEs assign a single vector per word, allowing us to directly inspect which words dominate each axis. We use Skip-Gram with Negative Sampling (SGNS) embeddings trained on time-separated corpora and apply Orthogonal Procrustes (Hamilton et al., 2016).[10] To compute the semantic change scores between time-separated embeddings, we use cosine similarity.

3.2 Results

RQ: Can SCDTour maintain interpretability with sorted/gathered *sense* axes? Figure 1 shows the accuracy on the WIT using three LLMs. Across all three models, SCDTour w/ the change-specific weight ($\lambda > 0$) performs comparably against ICA and SCDTour ($\lambda = 0$) for interpretability, confirming that our method maintains interpretability regardless of the underlying LLM. As shown by Yamagiwa et al. (2023), PCA fails to provide interpretable axes compared to ICA.

RQ: Can SCDTour solve the SCD task with the sorted/gathered axes? On the SCD ranking task,

[4] In this paper, we assume that the top-10 words in each axis serve as its representative words (Yamagiwa et al., 2023, 2024; Musil and Mareček, 2024).

[5] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

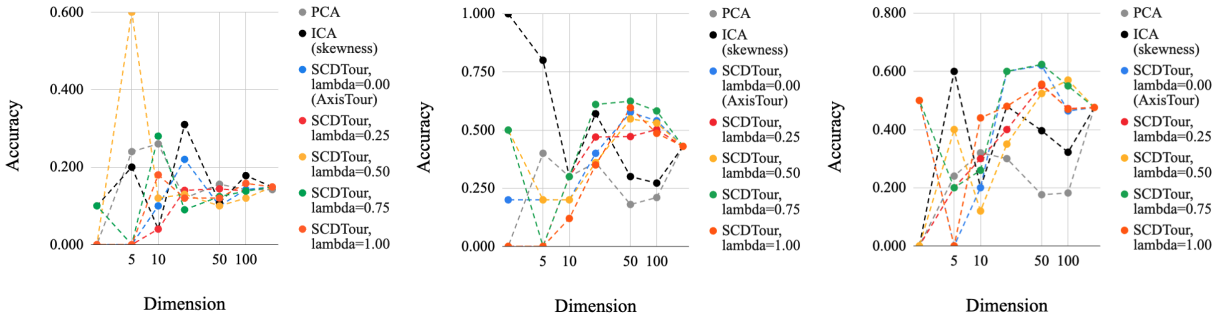
[6] https://huggingface.co/google/gemma-3-4b-it

[7] https://huggingface.co/Qwen/Qwen3-8B

[8] Dataset statistics are shown in §A.1.

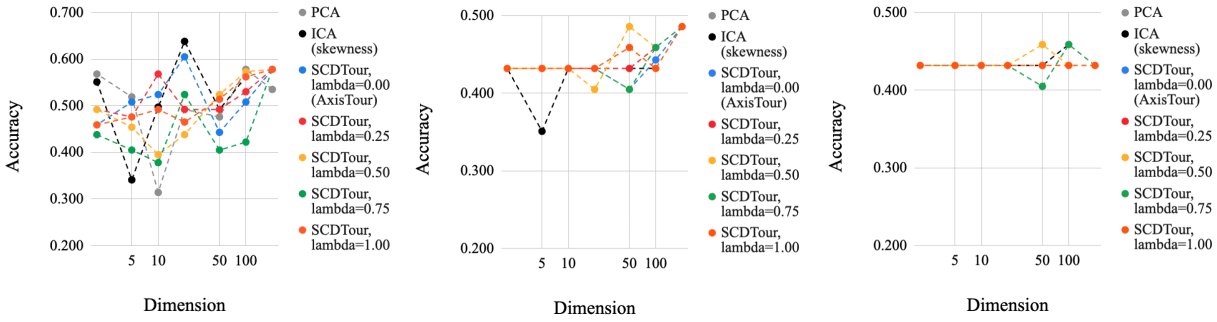
[9] Prompt templates for both WIT and binary SCD are shown in §A.3.

[10] We tune hyperparameters as described in §A.2.



(a) Llama-3.1 (b) Gemma-3 (c) Qwen3

Figure 1: Accuracy on the word intruder test.



(a) Llama-3.1 (b) Gemma-3 (c) Qwen3

Figure 2: Accuracy on the semantic change detection.

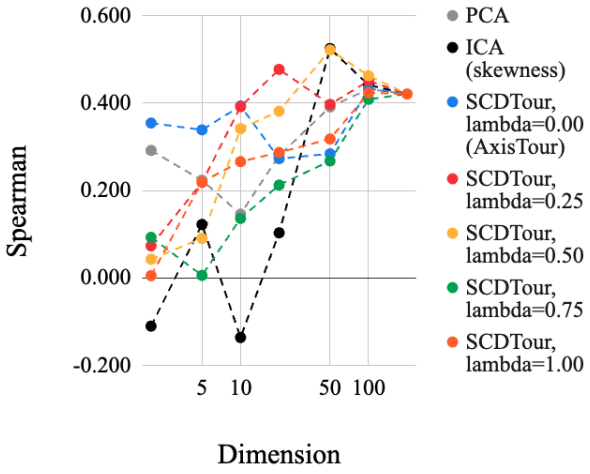


Figure 3: Spearman's rank correlation for the semantic change detection task.

Figure 3 shows that SCDTour ($\lambda = 0.25, 0.50$) outperforms baselines (PCA, ICA, and SCDTour ($\lambda = 0.00$)) in the low- and mid-dimensional settings ($k = 20, 50, 100$). While SCDTour ($\lambda = 0.00$) performs best in extremely low-dimensional settings ($k = 2, 5$), its interpretability is limited, as we discuss later. Figure 2 shows results across three LLMs. With Llama-3.1, SCDTour achieves the least degradation when reducing dimensionality, producing balanced judgments in few-shot settings. In contrast, Gemma-3 and Qwen3 frequently defaulted to NO outputs across most configurations, leading to reduced discrimination. This discrepancy may reflect differences in instruction-following capabilities.[11] Such differences in training strategy appear to strongly influence semantic judgment tasks that require alignment with human intuition (Sorensen et al., 2022).

To further investigate the interpretability of the learned axes, we analyse two representative target words: *graft*, which underwent a semantic change (from *horticultural grafting* to *medical transplant*), and *chairman*, which maintained a stable meaning over time. Results are shown in Table 3. We see that at $d = 200$ (full dimension) and $k = 100$, most methods can retrieve interpretable axes whose top-ranked words correspond to the meanings in each time period. In contrast, PCA constantly fails to extract such axes even at $k = 100$, highlighting its limited utility for interpretability. For $k = 20$, SCDTour ($\lambda = 0.25$) successfully identifies axes that capture each relevant meaning of the target words. However, SCDTour ($\lambda = 0$) fails to capture

[11] LLaMA-3.1 has been fine-tuned using human feedback data, whereas the extent to which Gemma-3 and Qwen3 rely on synthetic/human-generated instruction data remains unclear.

Method	word: *graft*		word: *chairman*	
	axis	words	axis	words
		$d = 200$ (full)		
Raw	70	**flowering**, **pear**, **pendulous**, **deciduous**, **sycamore**	42	**vote**, **election**, **senate**, **senator**, **delegate**
	157	**respiratory**, **intestinal**, **pulmonary**, **uterine**, **inflammation**	42	**caucus**, **senatorial**, **republican**, **democrat**, **nominee**
		$k = 100$		
PCA	65	conceivable, precaution, agatha, believe, alteration	94	globe, self-satisfied, isaiah, dainty, area
	66	prop, ragusa, malignity, thieving, vanegas	53	**poky**, avenue, t1, tint, omnipotence
ICA	70	**pear**, **sycamore**, **shrubbery**, **tree**, **fern**	99	+, –, ditto, sauk, hydrochloric
	82	realist, condensed, t', misapply, commodity	94	rid, **accuse**, dispose, whiff, incapable
SCDTour				
$\lambda = 0.00$	73	**tree**, **sycamore**, **pendulous**, **pear**, **deciduous**	17	vote, election, presidential, elect, candidate
	12	**inflammation**, **intestinal**, **pulmonary**, **respiratory**, **disease**	67	leo, **authority**, **department**, quarter, 20th
$\lambda = 0.25$	41	**pear**, **elm**, hampshire, **shrub**, **flowering**	15	**vote**, **election**, **senator**, **senate**, **judiciary**
	92	**pulmonary**, **respiratory**, **infection**, **colon**, **liver**	15	**republican**, **caucus**, **congressional**, **senatorial**, **gubernatorial**
$\lambda = 1.00$	5	**flowering**, **sycamore**, **shrub**, **pear**, **herbaceous**	53	**vote**, **senator**, **election**, **representative**, **ballot**
	80	givin, intestinal, seein, **pulmonary**, **respiratory**	12	incapable, dozens, **accuse**, fond, kind
		$k = 20$		
PCA	14	hutchinson, montague, rain, mosquito, kirk	10	disbelief, god, almighty, forgiveness, whosoever
	14	louisiana, piping, lilac, predominate, starch	14	louisiana, piping, lilac, predominate, starch
ICA	11	exempt, extricate, exemption, deviate, detract	14	address, customary, ducat, color, alice
	13	beyond, shirt, **medication**, determine, revolution	17	cease, lately, connection, lydia, already
SCDTour				
$\lambda = 0.00$	13	amends, precedence, necessary, precaution, observation	3	**indictment**, **politician**, **lunatic**, **defendant**, **adjudge**
	2	**transplant**, **inflammation**, **infection**, **disorder**, **respiratory**	2	transplant, inflammation, infection, disorder, respiratory
$\lambda = 0.25$	8	**pear**, **vine**, **tree**, **elm**, **pendulous**	3	**magistracy**, **elect**, curtis, **amendment**, **legislature**
	18	**respiratory**, chronic, **infection**, **pulmonary**, **renal**	3	**assembly**, **congressional**, driver, **nominee**, **caucus**
$\lambda = 1.00$	1	hemlock, stunted, moneywort, **crop**, **apple-tree**	10	artillery, picket, apartment, palace, portico
	0	2,200, audible, sidle, syllable, difference	2	fond, **accuse**, let, plenty, faster
		$k = 5$		
PCA	4	pie, mince, first-rate, stuff, tight	3	glasgow, 1835, 43, sloop, 1830
	4	tasty, **pill**, prescription, **medication**, dessert	3	kentucky, oakland, 153, md, fl
ICA	0	hebraic, ludicrousness, orvieto, tannin, wattie	3	qui, je, vous, comme, zo
	0	gainsay, hoyden, condi, monarchial, bb	3	sus, que, por, la, como
SCDTour				
$\lambda = 0.00$	3	militia, thornton, assistance, impression, opportunity	0	1794, roldan, misma, -, ce
	3	enormous, specimen, revelation, outstreched, handkerchief	0	ciudad, mrs, cell, marietta, **breed**
$\lambda = 0.25$	2	laboratory, deduce, imaginary, varmint, jury	2	laboratory, deduce, imaginary, varmint, jury
	4	**illness**, **dental**, m, nightgown, hem	2	clearly, greased, cultivated, legitimate, hamlet
$\lambda = 1.00$	3	cambridge, xx_y, aime, mocha, tut	2	billows, manufactory, prevail, taste, jerry
	0	unison, credible, anticipated, $800, dramatically	0	unison, credible, anticipated, $800, dramatically

Table 3: Representative words from the most activated axis of the target word embedding (*graft* and *chairman*) at two time periods t_1 (shown in gray) and t_2 (in black), across different methods. For each method and target word, we identify the axis with the highest value in the embedding, and list the top-5 words associated with that axis. Words that reflect the meaning of the target word are highlighted in **bold**.

the axis representing the older meaning of *graft*. When the number of dimensions is reduced to 5, no method reliably produces axes with coherent representative words. In such cases, the top words on each axis lack semantic consistency and fail to reflect interpretable meanings. We hypothesise that this is because the number of dimensions becomes insufficient to encode the full range of word meanings, causing multiple unrelated axes to be merged, according to Equation 3. A similar trend is observed for the qualitative analysis in Appendix B.

Overall, these findings demonstrate that the use of the change-specific weight enables SCDTour to efficiently maintain interpretability and performance, even with a reduced number of dimensions.

4 Conclusion

We presented SCDTour, a method that orders and merges interpretable axes using meaning- and change-specific weights. It maintains interpretability while preserving SCD performance, even in low-dimensional settings. In future work, we plan to extend the comparison to contextualised embeddings and multilingual SCD tasks, in order to more broadly evaluate the trade-offs between interpretability and performance.

Limitations

While our proposed method, SCDTour, demonstrates promising results, it has the following limitations.

First, our evaluations are conducted only for En-

glish, which is a morphologically limited language. This is due to the necessity of both quantitative evaluation and qualitative analysis, which require extensive lexical resources and contextual understanding that are readily available for English. However, our proposed method is language-agnostic, and we expect that it would generalise to languages other than English.

Second, we focus exclusively on SWEs in this paper. As discussed in §3.1, SWEs provide a single vector per word, enabling direct inspection of axis-level information, and has been adopted in prior work (Sato, 2022; Yamagiwa et al., 2024; Musil and Mareček, 2024). In contrast, CWEs encode multiple sense-aware vectors per token, making axis-level interpretation more challenging. While we prioritised interpretability in our analysis, CWE often offers stronger performance in SCD. A recent study has shown that SCD-specific dimensions exist in CWE (Aida and Bollegala, 2025), suggesting that future work could extend SCDTour to contextualised embeddings.

Ethical Considerations

This paper does not introduce new datasets or models. We conduct our experiments using existing datasets and pre-trained models. To the best of our knowledge, no ethical issues have been reported regarding those evaluation datasets (SemEval-2020 Task 1 English (Schlechtweg et al., 2020), derived from CCOHA (Alatrash et al., 2020)). Pre-trained models, such as GloVe and LLama-3.1, may contain social biases (Kaneko and Bollegala, 2019; Basta et al., 2019; Oba et al., 2024). Future work should assess how these biases might be reflected in the obtained axes and affect interpretation in real-world applications.

Acknowledgements

Taichi Aida would like to acknowledge the support by JST, the establishment of university fellowships towards the creation of science technology innovation, Grant Number JPMJFS2139.

References

Taichi Aida and Danushka Bollegala. 2023a. Swap and predict – predicting the semantic changes in words across corpora by context swapping. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7753–7772, Singapore. Association for Computational Linguistics.

Taichi Aida and Danushka Bollegala. 2023b. 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 and Danushka Bollegala. 2024. A semantic distance metric learning approach for lexical semantic change detection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7570–7584, Bangkok, Thailand. Association for Computational Linguistics.

Taichi Aida and Danushka Bollegala. 2025. Investigating the contextualised word embedding dimensions specified for contextual and temporal semantic changes. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1413–1437, Abu Dhabi, UAE. 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 Lingustics.

Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2020. CCOHA: Clean corpus of historical American English. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6958–6966, Marseille, France. European Language Resources Association.

Naomi Baes, Nick Haslam, and Ekaterina Vylomova. 2024. A multidimensional framework for evaluating lexical semantic change with social science applications. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1390–1415, Bangkok, Thailand. Association for Computational Linguistics.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy. Association for Computational Linguistics.

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.

Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of the 57th Annual*

Meeting of the Association for Computational Linguistics, pages 457–470, Florence, Italy. Association for Computational Linguistics.

Mario Giulianelli, Iris Luden, Raquel Fernandez, and Andrey Kutuzov. 2023. Interpretable word sense representations via definition generation: The case of semantic change analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3130–3148, Toronto, Canada. 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.

Keld Helsgaun. 2000. An effective implementation of the lin–kernighan traveling salesman heuristic. *European Journal of Operational Research*, 126(1):106–130.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1641–1650, Florence, Italy. Association for Computational Linguistics.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA. Association for Computational Linguistics.

Andrey Kutuzov, Mariia Fedorova, Dominik Schlechtweg, and Nikolay Arefyev. 2024. Enriching word usage graphs with cluster definitions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6189–6198, Torino, Italia. ELRA and ICCL.

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, Lilja Ovrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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.

Xianghe Ma, Michael Strube, and Wei Zhao. 2024. Graph-based clustering for detecting semantic change across time and languages. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1542–1561, St. Julian's, Malta. Association for Computational Linguistics.

Tomáš Musil and David Mareček. 2024. Exploring interpretability of independent components of word embeddings with automated word intruder test. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6922–6928, Torino, Italia. ELRA and ICCL.

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.

Daisuke Oba, Masahiro Kaneko, and Danushka Bollegala. 2024. In-contextual gender bias suppression for large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1722–1742, St. Julian's, Malta. Association for Computational Linguistics.

Francesco Periti, David Alfter, and Nina Tahmasebi. 2024a. Automatically generated definitions and their utility for modeling word meaning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14008–14026, Miami, Florida, USA. Association for Computational Linguistics.

Francesco Periti, Pierluigi Cassotti, Haim Dubossarsky, and Nina Tahmasebi. 2024b. Analyzing semantic change through lexical replacements. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4495–4510, Bangkok, Thailand. Association for Computational Linguistics.

Guy D. Rosin, Ido Guy, and Kira Radinsky. 2022. Time masking for temporal language models. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, pages 833–841, New York, NY, USA. Association for Computing Machinery.

Guy D. Rosin and Kira Radinsky. 2022. Temporal attention for language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1498–1508, Seattle, United States. Association for Computational Linguistics.

Ryoma Sato. 2022. Word tour: One-dimensional word embeddings via the traveling salesman problem. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2166–2172, Seattle, United States. 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.

Taylor Sorensen, Joshua Robinson, Christopher Rytting, Alexander Shaw, Kyle Rogers, Alexia Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An information-theoretic approach to prompt engineering without ground truth labels. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 819–862, Dublin, Ireland. Association for Computational Linguistics.

Xiaohang Tang, Yi Zhou, Taichi Aida, Procheta Sen, and Danushka Bollegala. 2023. Can word sense distribution detect semantic changes of words? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3575–3590, Singapore. Association for Computational Linguistics.

Hiroaki Yamagiwa, Momose Oyama, and Hidetoshi Shimodaira. 2023. Discovering universal geometry in embeddings with ICA. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4647–4675, Singapore. Association for Computational Linguistics.

Hiroaki Yamagiwa, Yusuke Takase, and Hidetoshi Shimodaira. 2024. Axis tour: Word tour determines the order of axes in ICA-transformed embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 477–506, Miami, Florida, USA. Association for Computational Linguistics.

Zhao yu Su, Zecheng Tang, Xinyan Guan, Juntao Li, Lijun Wu, and M. Zhang. 2022. Improving temporal generalization of pre-trained language models with lexical semantic change. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6380–6393, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Experimental Details

A.1 Data Statistics

To evaluate the effectiveness of the SCDTour, we performed a preliminary experiment using the Word Embedding Benchmarks, which include three subtasks: categorical, similarity, and analogy. The benchmark is publicly available under the MIT License.[12]

In the SemEval-2020 Task 1 for English, the dataset is constructed from the Cleaned Corpus of Historical American English (CCOHA) (Alatrash et al., 2020), which contains time-separated newspapers, magazines, and (non-)fiction books suitable for analysing diachronic semantic change.[13] Table 4 summarises the statistics of the lemmatised version of the corpora, including time periods, the number of target words and tokens.

Time Period	#Targets	#Tokens
1810s–1860s	37	6.5M
1960s–2010s		6.7M

Table 4: Statistics of the SCD benchmark, SemEval-2020 Task 1 (English). #Targets and #Tokens indicate the number of target words and tokens.

A.2 Hyperparameters

We used the pre-trained `GloVe 6B`[14] in the preliminary experiment. For the main SCD experiments, which aim to investigate how performance and interpretability can be maintained even after dimension reduction, we follow the procedure of Laicher et al. (2021) for selecting hyperparameters for the SGNS model. Specifically, we perform a grid search over the values in Table 5.[15] For each configuration, we evaluate the model performance on the SCD ranking task and select the best setting that yields the highest Spearman's correlation. This hyperparameter tuning is performed jointly across different time periods. Table 5 also shows the best setting, which is used in the main experiments presented in §3.2.

[12] https://github.com/kudkudak/word-embeddings-benchmarks

[13] This dataset is licensed under the Creative Commons Attribution 4.0 International License.

[14] It is available at https://nlp.stanford.edu/projects/glove/ under the Public Domain Dedication and License.

[15] We used the LSCDetection toolkit available at https://github.com/Garrafao/LSCDetection . It is licensed under the GNU General Public License.

Parameter	Values
window size	5, **10**
dimension	50, 100, **200**, 300
iteration	**5**, 10, 20, 30
negative samples	5

Table 5: Hyperparameters used for SGNS. **Bold** values indicates the best settings.

Word Intruder Test

Which word does not fit the following group of words?
{*Top-4 words in i-th axis and one random word from different axis*}

Answer strictly using just one word.
Do not provide any additional explanation.

Figure 4: Prompt used for word intruder test. This prompt is referred to Musil and Mareček (2024).

A.3 Prompts

In the WIT and SCD binary task, we report the average accuracy over five runs to account for any variance in the generation process using three models: `Llama-3.1-8B-Instruct`, `Gemma-3-4b-it`, and `Qwen3-8B`[16]. Prompts for each task are shown in Figure 4 and Figure 5.

B Additional SCD Results

In addition to the analysis presented in § 3, we provide further case studies on two word pairs: (i) *plane* (shifted from *(mathematical) surface* to *aircraft*) and *tree* (stable), and (ii) *attack* (extended to include *heart attack*) and *relationship* (stable). Table 6 and Table 7 confirm that SCDTour can preserve the interpretability of semantic change while the dimension is reduced to $k = 20$. We observe the following trends across all target words described in § 3. At $k = 100$, all methods can retrieve representative words corresponding to the meaning of the target word. SCDTour maintains the ability to extract the relevant words at $k = 20$, demonstrating robustness against baselines. However, at $k = 5$, no method can correctly capture the meaning of the target word, due to excessive merging of distinct meanings into a single axis.

[16]These models are licensed under the Llama 3.1 Community License, Gemma License, and Apache 2.0 License, respectively.

Semantic Change Detection

You are given two sets of words where each set collectively conveys a particular meaning. We use standard set notation to represent a set, where words are separated by commas.
If Set 1 and Set 2 both express the same meaning, answer YES. Otherwise answer NO.
Examples:
Set 1 = {jujube, peach, plum, olive, cherry}
Set 2 = {melon, banana, apple, mango, berry}
Answer: YES

Set 1 = {scarf, skirt, trouser, suit, shirt}
Set 2 = {supercomputer, multiprocessor, file server, personal computer, minicomputer}
Answer: NO

Below are two sets of words for evaluation.

Set 1 = {*Top-5 words most responsive to the principal axis of word embedding* $\mathbf{e}^w_{t_1}$.}
Set 2 = {*Top-5 words most responsive to the principal axis of word embedding* $\mathbf{e}^w_{t_2}$.}
Answer:

Figure 5: Prompt used for semantic change detection (binary classification).

C Prompt Design and LLM Behaviour

While zero-shot prompts worked well for WIT, they did not yield meaningful outputs in the SCD binary classification task: the LLM consistently returned YES or NO regardless of the input sets. To address this issue, and following Sorensen et al. (2022), we designed few-shot prompts using WordNet[17] synsets. We constructed both strong positive and negative examples from these synsets, as shown in Figure 5. We found that these few-shot prompts help the LLM toward recognising subtle semantic differences between given sets of words, mitigating their strict similarity thresholds observed in Sorensen et al. (2022).

In addition to the prompting strategy, we observed that both instruction tuning and explicit instructions were essential for the WIT and SCD. Without instruction-tuned models such as `meta-llama/Llama-3.1-8B`, the LLM often failed to follow the task setup. Even with instruction-tuned models, ambiguous prompts lead LLMs to return generic outputs (e.g. "Sure!"). To mitigate this issue, we included clear phrases (e.g. "Answer strictly using just one word."), which improved consistency.

[17]`https://www.nltk.org/howto/wordnet.html`

Method	word: *plane*		word: *tree*	
	axis	words	axis	words
		$d = 200$ (full)		
Raw	28	**z**, **q**, **g**, **k**, **h**	70	**flowering**, **pear**, **pendulous**, **deciduous**, **sycamore**
	92	**ship**, **aboard**, **crew**, **reconnaissance**, **passenger**	70	**shrub**, **beech**, **tree**, **leaf**, **dogwood**
		$k = 100$		
PCA	99	1200, not, oakwood, lie, discovery	73	**tulip**, gouverneur, dignity, dreamy, sive
	90	perpetrate, honorably, forlom, somber, **voyager**	73	$60, garibaldi, basilikon, winter, publication
ICA	70	pear, sycamore, shrubbery, tree, fern	70	**pear**, **sycamore**, **shrubbery**, **tree**, **fern**
	92	**ship**, **aboard**, uss, **passenger**, **reconnaissance**	70	**shrub**, **beech**, **cactus**, **tree**, **maple**
SCDTour				
$\lambda = 0.00$	43	**z**, **h**, **g**, **q**, **ky**	73	**tree**, **sycamore**, **pendulous**, **pear**, **deciduous**
	56	**ship**, **aboard**, **passenger**, **crew**, **air**	73	**shrub**, **leaf**, **beech**, **cactus**, **tree**
$\lambda = 0.25$	80	**z**, **abc**, **h**, **g**, **q**	41	**pear**, **elm**, hampshire, **shrub**, **flowering**
	68	**ship**, **crew**, **aboard**, **passenger**, uss	41	hampshire, broome, rochelle, powel, orleans
$\lambda = 1.00$	64	**z**, **abc**, **g**, **h**, **ky**	5	**flowering**, **sycamore**, **shrub**, **pear**, **herbaceous**
	13	**ship**, **aboard**, **flight**, uss, **crew**	5	**shrub**, **beech**, **cactus**, **deciduous**, **dogwood**
		$k = 20$		
PCA	10	disbelief, god, almighty, forgiveness, whosoever	15	turner, protege, robertson, kidnapper, memphis
	16	sell, buy, bourse, outfit, fancy	13	defendant, squalid, humbert, detention, prison
ICA	10	utterly, pour, word, vegetation, russia	11	exempt, extricate, exemption, deviate, detract
	14	7,000, gordon, dip, magareta, edge	11	derive, exempt, detract, emanate, refrain
SCDTour				
$\lambda = 0.00$	3	indictment, politician, lunatic, defendant, adjudge	14	guardianship, **twig**, **oak**, **bough**, **flowering**
	11	delighted, merge, sob, equip, inclined	14	owl, squirrel, assortment, **tree**, **conservation**
$\lambda = 0.25$	16	**z**, **q**, **w**, **g**, **arc**	8	**pear**, **vine**, **tree**, **elm**, **pendulous**
	13	earnings, liquidate, **regulator**, depositor, underwrite	8	broome, hampshire, mistake, soon, **fern**
$\lambda = 1.00$	12	**z**, **g**, **ky**, **q**, **sm**	1	**hemlock**, stunted, moneywort, **crop**, **apple-tree**
	9	**drawbridge**, progress, **aerial**, lu, episode	1	earnings, depositor, **shrub**, dividend, seller
		$k = 5$		
PCA	4	pie, mince, first-rate, stuff, tight	3	glasgow, 1835, 43, sloop, 1830
	3	kentucky, oakland, 153, md, fl	3	kentucky, oakland, 153, md, fl
ICA	3	qui, je, vous, come, zo	2	vanity, emotion, apprehension, temporary, petty
	2	joy, anger, disappointment, anxiety, sorrow	3	sus, queue, por, la, como
SCDTour				
$\lambda = 0.00$	2	spec, r, ben, heave, brimstone	3	militia, thornton, assistance, impression, opportunity
	2	sussex, premature, deck, 262, grunt	3	enormous, specimen, revelation, outstretched, handkerchief
$\lambda = 0.25$	0	consistent, lunatic, facility, pikes, unacquainted	2	laboratory, deduce, imaginary, varmint, jury
	3	lecture, oatmeal, dark, objective, industry	2	clearly, greased, **cultivated**, legitimate, hamlet
$\lambda = 1.00$	3	cambridge, xx_v, aimed, mocha, tut	0	web, whereby, resounding, bind, last
	3	mystery, biblical, z, dairy, fais	0	unison, credible, anticipated, $800, dramatically

Table 6: Representative words from the most activated axis of the target word embedding (*plane* and *tree*) at two time periods t_1 (shown in gray) and t_2 (in black), across different methods. For each method and target word, we identify the axis with the highest value in the embedding, and list the top-5 words associated with that axis. Words that reflect the meaning of the target word are highlighted in **bold**. *Plane* underwent a semantic change (*(mathematical) surface* to *aircraft*), while *tree* remained stable. At $d = 200$ and $k = 100$, all methods capture the corresponding meanings for both time periods. SCDTour ($\lambda = 0.25$) maintains this interpretability even at $k = 20$, whereas other methods such as PCA or ICA struggle. At $k = 5$, no method successfully preserves corresponding meanings due to the limited number of axes or excessive merging.

Method	word: *attack*		word: *relationship*	
	axis	words	axis	words
		$d = 200$ (full)		
Raw	49	amends, **arrangement**, **debut**, **effort**, **appearance**	34	**aryan**, **discrepancy**, clavicle, **demarcation**, **estrangement**
	157	**respiratory**, **intestinal**, **pulmonary**, **uterine**, **inflammation**	34	**correlation**, **disparity**, **relationship**, **discrepancy**, **distinction**
		$k = 100$		
PCA	93	**alert**, **fated**, seventh, ihe, **resentment**	94	globe, self-satisfied, isaiah, dainty, **area**
	77	declaration, **throbbing**, holland, i'he, weatherworn	69	**friendship**, bridge, **bond**, ingly, self-denying
ICA	70	pear, sycamore, shrubbery, tree, fern	51	**fellow-man**, protege, pursuer, townsman, fellow-citizen
	54	minister, prime, beautiful, mountain, marriage	71	accordance, **sympathize**, **interfere**, **coincide**, **comply**
SCDTour				
$\lambda = 0.00$	90	rely, dependent, **encroach**, devolve, **preye**	19	**aryan**, **russia**, **discrepancy**, **austria**, **sweden**
	90	embark, **verge**, rely, depending, reliance	19	**austria**, **scandinavia**, **belgium**, **albania**, **ethiopia**
$\lambda = 0.25$	42	**arrangement**, amends, **appearance**, confession, **debut**	57	**aryan**, **discrepancy**, **russia**, clavicle, **scandinavian**
	92	**pulmonary**, **respiratory**, **infection**, **colon**, **liver**	57	**austria**, **scandinavia**, **belgium**, **albania**, 1715
$\lambda = 1.00$	52	**artillery**, **cavalry**, **regiment**, **troop**, **infantry**	2	shrill, peal, reverberate, tinkling, dirge
	22	embark, dote, depending, depend, reliance	2	shrill, **correlation**, cymbal, muffle, hoarse
		$k = 20$		
PCA	12	**camp**, appetite, swallow, draught, pipe	19	**banter**, data, **affinity**, **disinclination**, meaning
	11	polka, profane, trickster, wilde, gobbler	18	underlie, **collapse**, sweaty, na, ta
ICA	12	god, cook, thus, throw, convincing	10	utterly, pour, word, vegetation, **russia**
	10	chain, atmosphere, interview, prove, phrase	10	chain, **atmosphere**, interview, prove, phrase
SCDTour				
$\lambda = 0.00$	14	**guardianship**, twig, oak, bough, flowering	10	neither, nor, plentitude, **neighbourhood**, **homelike**
	2	**transplant**, **inflammation**, **infection**, **disorder**, **respiratory**	3	winner, defendant, moslem, pennant, congressmen
$\lambda = 0.25$	14	d'etre, de, cet, **normandie**, rien	10	**abstain**, deduce, **exempt**, lately, alleghany
	14	pasado, muy, ramn, quelque, misma	11	**tension**, **belgium**, **confederation**, **unrest**, **albania**
$\lambda = 1.00$	4	ruminate, extremely, lavish, depending, preye	3	meantime, **assist**, meanwhile, inmost, reciprocate
	19	jamieson, jamie, galbraith, oliver, shrewsbury	0	2,200, audible, sidle, syllable, difference
		$k = 5$		
PCA	3	**glasgow**, 1835, 43, **sloop**, 1830	2	mortimer, digby, pauline, harding, terence
	3	Kentucky, oakland, 153, md, fl	4	tasty, pill, prescription, medication, dessert
ICA	3	qui, je, vous, come, zo	2	vanity, emotion, apprehension, temporary, petty
	2	joy, anger, disappointment, **anxiety**, sorrow	2	joy, anger, disappointment, anxiety, sorrow
SCDTour				
$\lambda = 0.00$	3	**militia**, thornton, **assistance**, impression, **opportunity**	2	spec, r, ben, heave, brimstone
	0	ciudad, mrs, cell, marietta, breed	0	ciudad, mrs, cell, marietta, breed
$\lambda = 0.25$	2	**laboratory**, deduce, imaginary, varmint, jury	2	laboratory, deduce, imaginary, **varmint**, jury
	0	gobble, chilton, consecutive, convenience, twenty-five	1	98, nonsense, perspective, havin, balk
$\lambda = 1.00$	1	conjugal, brightest, saturday, consider, greenish	0	web, whereby, resounding, bind, last
	4	gibson, brighten, colorless, coarse, ol	0	unison, credible, anticipated, $800, dramatically

Table 7: Representative words from the most activated axis of the target word embedding (*attack* and *relationship*) at two time periods t_1 (shown in gray) and t_2 (in black), across different methods. For each method and target word, we identify the axis with the highest value in the embedding, and list the top-5 words associated with that axis. Words that reflect the meaning of the target word are highlighted in **bold**. *Attack* exhibits a semantic shift through the inclusion of the medical sense (*heart attack*) in the later time period t_2, whereas *relationship* remains semantically stable. At $d = 200$ and $k = 100$, most methods correctly capture relevant meanings for both time periods. At $k = 20$, only SCDTour ($\lambda = 0.00, 0.25$) consistently identifies axes aligned with the new sense of *attack*, while other methods retrieve more unrelated terms. At $k = 5$, most methods fail to reflect either meaning due to the limited number of axes or excessive axis merging.