Statistical Semantic Change Detection via Usage Similarities

Taichi Aida
Hitotsubashi University
taichia@scl.sds.hit-u.ac.jp

Daichi Mochihashi
The Institute of Statistical Mathematics
daichi@ism.ac.jp

Hiroya Takamura
Artificial Intelligence Research Center, AIST
takamura.hiroya@aist.go.jp

Toshinobu Ogiso
National Institute for Japanese Language and Linguistics
togiso@ninjal.ac.jp

Mamoru Komachi
Graduate School of SDS
Hitotsubashi University
mamoru.komachi@r.hit-u.ac.jp

Abstract

Semantic change detection comprises two subtasks: **classification**, which predicts whether a target word has undergone a semantic shift, and **ranking**, which orders words according to the degree of their semantic change. While most prior studies concentrated on **ranking** subtask, the **classification** subtask plays an equally important role, since many practical scenarios require a yes/no decision on semantic change rather than a global ranking. In this work, we propose a novel statistical method that predicts the presence or absence of semantic change. While most existing approaches infer semantic change by comparing word embeddings across time periods or domains, our method directly models the diachronic/synchronic consistency of usage-level similarity scores. Our experiments on SemEval-2020 Task 1 and WUGS datasets demonstrate that the proposed formulation outperforms existing state-of-the-art embedding-based methods, and robustly detects semantic change across languages in both diachronic and synchronic settings.[1]

1 Introduction

The meanings of words naturally evolve over time and across domains. Detecting such semantic change is essential for linguistic and lexicographic research, as well as for studying cultural and societal dynamics (Traugott and Dasher, 2001; Cook and Stevenson, 2010). Beyond these humanities-oriented applications, recent work has highlighted the importance of semantic change detection for various additional purposes, including information retrieval (Kutuzov et al., 2018) and efficient updating of masked language models (Su et al., 2022).

The Semantic Change Detection (SCD) task aims to automatically identify words that have undergone semantic shift. Recent shared tasks such as SemEval-2020 Task 1 (Schlechtweg et al., 2020) and the WUGS (Schlechtweg et al., 2021) framework have established benchmark datasets and evaluation protocols for this task. There are two subtasks in SCD: **classification**, which predicts whether a target word is semantically changed, and **ranking**, which orders target words according to the degree of semantic change (Schlechtweg et al., 2020). Due to the unsupervised nature of the problem, most prior studies have focused on the ranking subtask, evaluating models based on similarity scores derived from word embeddings across two periods/domains (Rosin et al., 2022; Rosin and Radinsky, 2022; Cassotti et al., 2023; Periti and Tahmasebi, 2024; Aida and Bollegala, 2024). However, ranking-based evaluation has inherent limitations, including limited interpretability of raw similarity scores and uncertainty about which parts of the ranked list are reliable. These issues motivate a stronger emphasis on the classification subtask.

Method	EN	DE	LA	SV
SGNS (Rother et al., 2020)	73.0	54.2	45.0	61.3
SGNS (Pražák et al., 2020)	62.2	75.0	**70.0**	67.7
BERT (Asgari et al., 2020)	70.3	75.0	55.0	74.2
Pólya (**ours**)	**76.1**	**80.0**	N/A	**88.6**

Table 1: Accuracy (in %) in SemEval-2020 Task 1 (Schlechtweg et al., 2020); our method does not rely on word embeddings contrary to prior state-of-the-arts.

In this work, we propose a new method that statistically determines whether a word has undergone semantic change by assessing the diachronic/synchronic consistency of a set of usage-level similarity scores. We consider SCD under the assumption that usage-level similarity scores are available, as is the case in current benchmark datasets such as SemEval-2020 Task 1 and WUGS, where these scores are provided via human annotations. Rather than addressing how such similarity scores should be predicted, our focus is on how semantic change can be statistically determined once

[1]Source code is available at https://github.com/a1da4/usage-similarity-polya.

usage-level similarity information is given. This design choice allows us to study SCD independently of specific similarity estimation methods. To this end, our framework models these similarity distributions using the Pólya distribution, enabling us to test whether the scores from two time periods are likely to originate from the same underlying distribution (indicating semantic stability) or from distinct distributions (indicating semantic change).

2 Related Work

Both classification and ranking evaluations in the SemEval-2020 Task 1 (Schlechtweg et al., 2020) are derived from WUGS-style annotation graphs (Schlechtweg et al., 2021), which are constructed by collecting a fixed number of usage examples from two different time periods/domains for each target word. After that, human annotators rate pairs of usages on a four-point semantic similarity scale (Schlechtweg et al., 2018, 2024), and the weighted graph is processed to obtain both binary labels and continuous scores representing its degree of semantic change. These resources have provided a baseline for evaluating SCD systems.

Methods for SCD generally rely on comparing static (Kim et al., 2014; Kulkarni et al., 2015; Yao et al., 2018; Aida et al., 2021) or contextualized (Hu et al., 2019; Giulianelli et al., 2020; Rosin et al., 2022; Rosin and Radinsky, 2022; Cassotti et al., 2023; Aida and Bollegala, 2024) word embeddings. Despite these advances, existing methods focus on the ranking subtask, in part because similarity-based metrics naturally yield ordered scores. Therefore, open fundamental challenges in binary decision-making still remain.

3 Method

In this work, we also leverage the WUGS-style dataset described above. For a given word (e.g., *plane*), we collect N usages from texts in period A (e.g., '*in the horizontal plane*') and M usages from period B (e.g., '*the plane was in flight*').

We construct a matrix $\mathbf{X}$ of size $(N{+}M) \times (N{+}M)$ over all possible pairs of these $(N{+}M)$ usages, as shown in Figure 1, where some of its entries x_{ij} are given a similarity score 1 through 4 between usage i and usage j. We consider two cases with this matrix:

Without sense change. In this case, there is no distinction between the entries x_{ij} in $\mathbf{X}$ and thus all the annotated scores could be assumed to be

$$\mathbf{X} = \begin{array}{c} \\ 1 \\ \\ N \\ N{+}1 \\ \\ N{+}M \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & & & & N & N{+}1 & & N{+}M \end{array} \\ \left(\begin{array}{cccc|cccc} 4 & & & & & 2 & & \\ & 4 & 3 & & & & 3 & 4 \\ & 2 & & & 2 & & & \\ 1 & & & & 2 & & & \\ \hline 1 & & & 4 & & 4 & 1 & \\ & & 2 & & & & & \\ 2 & & & & 2 & 3 & & \\ 3 & & & & & & & 4 \end{array}\right) \end{array}$$

Figure 1: Example of the score matrix $\mathbf{X}$ of usage similarities. Some of the usage pairs are annotated to have similarity scores, 1–4 in this case.

generated from the same underlying distribution $\mathbf{p}=(p_1,p_2,p_3,p_4)$ over the scores 1 through 4:

$$x_{ij} \sim \mathbf{p}_0 \quad \text{i.i.d.} \quad (1 \le i,j \le N{+}M) \tag{1}$$

With sense change. If there is a semantic change between the periods A and B, the similarity score over the pairs of usages within the same period or between the different periods will differ. Therefore, entry x_{ij} is assumed to generated from one of the four distinct distributions $\mathbf{p}_n$ $(n=1,\cdots,4)$ for the associated block of $\mathbf{X}$ delineated in Figure 1.

$$x_{ij} \sim \begin{cases} \mathbf{p}_1 & (1 \le i,j \le N) \\ \mathbf{p}_2 & (1 \le i \le N, N{+}1 \le j \le N{+}M) \\ \mathbf{p}_3 & (N{+}1 \le i \le N{+}M, 1 \le j \le N) \\ \mathbf{p}_4 & (N{+}1 \le i,j \le N{+}M) \end{cases} \tag{2}$$

As a generative model we need a prior for $\mathbf{p}=\mathbf{p}_0,\mathbf{p}_1,\cdots,\mathbf{p}_4$, and the simplest choice is a Dirichlet distribution

$$p(\mathbf{p}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} p_k^{\alpha_k - 1} \tag{3}$$

where $K=4$ in our case and $\boldsymbol{\alpha}=(\alpha_1,\cdots,\alpha_K)$ is a hyperparameter for this prior distribution. We employed $\boldsymbol{\alpha}=(1,\cdots,1)$, *i.e.*, uniform distribution over probability simplex, throughout this study.

Let a binary latent variable θ denote whether there is no semantic change ($\theta=0$) or there is a change ($\theta=1$). Then the likelihood of the data $\mathbf{X}$ when $\theta=0$ is given as follows:

$$\begin{aligned} p(\mathbf{X}|\theta{=}0) &= p(\mathbf{X}) = \int p(\mathbf{X}|\mathbf{p})p(\mathbf{p})d\mathbf{p} \\ &= \int \prod_{i,j} \prod_{k=1}^{K} p_k^{\mathbb{I}(x_{ij}=k)} \cdot \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} p_k^{\alpha_k-1} d\mathbf{p} \\ &= \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(L{+}\sum_k \alpha_k)} \prod_{k=1}^{K} \frac{\Gamma(\alpha_k{+}n_k)}{\Gamma(\alpha_k)} \end{aligned} \tag{4}$$

where $n_k = \sum_{ij} \mathbb{I}(x_{ij} = k)$ is the frequency of score k in $\mathbf{X}$ and L is the number of annotated entries.

This formula (4) is known as a Pólya distribution (Minka, 2000; Murphy, 2022). For the case semantic change, the likelihood is a product of block-wise Pólya distributions according to the Equation (2):

$$p(\mathbf{X}|\theta=1) = \prod_{n=1}^{4} p(\mathbf{X}_n) \tag{5}$$

Here, each $\mathbf{X}_n$ $(n=1, \cdots, 4)$ is a submatrix of $\mathbf{X}$ defined by Equation (2), and n_k and L are similarly computed within each $\mathbf{X}_n$. Therefore, when $p(\theta=0) = p(\theta=1) = 1/2$, we can compute a posterior probability of θ as follows:

$$p(\theta|\mathbf{X}) \propto p(\mathbf{X}|\theta)p(\theta) \propto \begin{cases} p(\mathbf{X}|\theta=0) \\ p(\mathbf{X}|\theta=1) \end{cases} \tag{6}$$

where $p(\mathbf{X}|\theta=0)$ and $p(\mathbf{X}|\theta=1)$ are given by Equations (4) and (5), respectively. Intuitively speaking, this probability measures whether the observed score matrix is homogeneous or not.

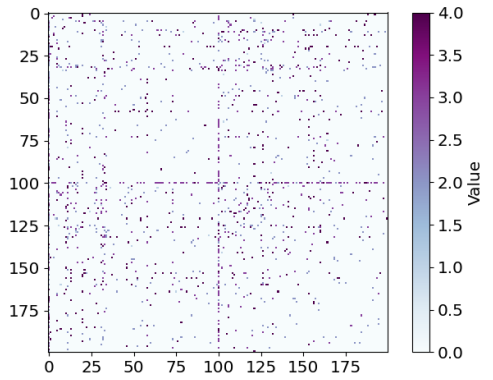
4 Experiments

Datasets We evaluate our method on two benchmark resources: SemEval-2020 Task 1 and a subset of the WUGS datasets.[2]

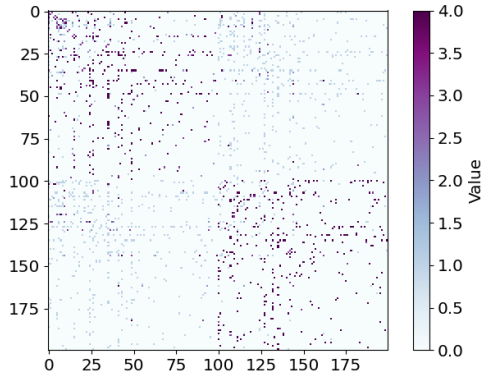
Evaluation We evaluate all methods using Accuracy, following the standard classification protocol adopted in the WUGS framework. Each target word is labeled as stable or changed, and predictions are compared to gold-standard WUGS-derived binary labels.

Baselines For SemEval-2020 Task 1, we compare our proposal against the three best-performing systems for each language reported in the shared task (Rother et al., 2020; Pražák et al., 2020; Asgari et al., 2020). All of these baselines rely on static or contextualized word embeddings. For WUGS datasets, where system outputs are not directly comparable across languages or domains, we adopt a simple baseline: MostFreq, which predicts the majority class (stable or changed) for each dataset.

Proposed Method To provide intuition for our method in practice, we illustrate how the proposed



(a) "*Attack*"



(b) "*Plane*"

Figure 2: Visualization of score matrices for *attack* and *plane*. *Attack* represents a word whose meaning has not changed, whereas *plane* represents a word that has undergone semantic change. In each graph, instances indexed from 0–99 correspond to the earlier period, and those from 100–199 correspond to the later period. Annotations are assigned on a scale from 1 (Unrelated) to 4 (Identical), while 0 indicates no annotation or an unknown label (Schlechtweg et al., 2021).

decision rule behaves on individual targets. Figure 2 shows the score matrices for *attack* (a semantically stable word) and *plane* (a semantically changed word).[3] This figure highlights a key intuition behind our method: for semantically stable/changed words, the score matrix tends to be homogeneous/heterogeneous. The prediction is made by comparing the log-likelihoods under the **Stable** and **Changed** hypotheses (we predict **Changed** when $p(\mathbf{X}|\theta=1) > p(\mathbf{X}|\theta=0)$). For *attack*, we obtain $\log p(\mathbf{X}|\theta=0) = -6213.9$ and $\log p(\mathbf{X}|\theta=1) = -6282.8$, so the model favors the **Stable** hypothesis; the gold label is **Stable**. For

[2]Datasets are available at https://www.ims.uni-stuttgart.de/en/research/resources/experiment-data/wugs/. We select only datasets with usage-usage similarity annotations, since our method requires similarity-score distributions.

[3]The cross-shaped pattern in Figure 2a reflects the WUGS annotation procedure, which prioritizes informative usage pairs rather than annotating all pairs exhaustively (Schlechtweg et al., 2021). As a result, some usages are compared with many others, while many entries remain unannotated (0), producing the observed cross-shaped structure.

Data	Language	Grouping 1	Grouping 2	Accuracy (%) MostFreq	Pólya
DWUG EN	EN	1810–1860	1960–2010	54.3	**76.1**
DWUG EN Resampled	EN	1810–1860	1960–2010	60.0	**80.0**
DWUG DE	DE	1800–1899	1946–1990	60.0	**80.0**
DWUG DE Resampled	DE	1800–1899	1946–1990	60.0	**73.3**
DiscoWUG	DE	1800–1899	1946–1990	51.0	**72.0**
RefWUG	DE	1750–1800	1850–1900	**54.5**	45.5
DURel	DE	1750–1800	1850–1900	**63.6**	**63.6**
SURel	DE	general	domain specific	63.6	**68.2**
RuSemShift 1	RU	1682–1916	1918–1990	**77.5**	**77.5**
RuSemShift 2	RU	1918–1990	1991–2016	**62.3**	**62.3**
RuShiftEval 1	RU	1700–1916	1918–1990	**74.8**	**74.8**
RuShiftEval 2	RU	1918–1990	1992–2016	**70.3**	**70.3**
RuShiftEval 3	RU	1700–1916	1992–2016	**68.5**	**68.5**
DWUG ES	ES	1810–1906	1994–2020	55.5	**78.0**
DiaWUG	ES	Spanish variant 1	Spanish variant 2	65.6	**81.3**
DWUG SV	SV	1790–1830	1895–1903	68.1	**88.6**
DWUG SV Resampled	SV	1790–1830	1895–1903	60.0	**73.3**
ChiWUG	ZH	1954–1978	1979–2003	**57.5**	52.5
DWUG IT	IT	1948–1970	1990–2014	**69.2**	N/A
DWUG LA	LA	−200–0	0–2000	**55.5**	N/A
NorDiaChange 1	NO	1929–1965	1970–2013	67.5	**75.0**
NorDiaChange 2	NO	1980–1990	2012–2019	**77.5**	70.0

Table 2: Results on WUGS datasets. For each dataset, we report accuracy for the MostFreq baseline and our Pólya-based method. The proposed approach yields strong performance across languages and dataset types, demonstrating robustness in both diachronic and synchronic SCD.

plane, we obtain $\log p(\mathbf{X}|\theta=0) = -7427.2$ and $\log p(\mathbf{X}|\theta=1) = -7095.6$, leading the model to favor the **Changed** hypothesis; the gold label is **Changed**. These examples illustrate the qualitative behavior of our decision rule and provide intuition for the quantitative results presented next.

Results Table 1 summarizes the results on SemEval-2020 Task 1. Despite not relying on any word embeddings, our method achieves higher accuracy than all embedding-based state-of-the-art systems in three out of four languages, demonstrating the effectiveness of modeling temporal consistency in usage similarity distributions. To assess the robustness of our framework beyond the SemEval setting, Table 2 presents results on the WUGS datasets. Our method consistently attains strong accuracy across all languages and performs robustly in both diachronic and synchronic settings. These findings indicate that the proposed framework generalizes well beyond the SemEval datasets and is applicable to diverse languages and domains.

Discussion and Future Work Although our experiments rely on human usage-similarity annotations, recent studies suggest that high-quality labels can be obtained easily through automatic methods. The DURel Annotation Tool (Schlechtweg et al., 2024) already integrates XL-LEXEME (Cassotti et al., 2023) to automatically propose usage-similarity annotations. Moreover, Periti and Tahmasebi (2024) show that contextualized embedding models and large language models can generate annotation labels that approach the quality of human annotations. These developments indicate that our statistical framework may be deployed in a fully automatic setting in the future, where usage-similarity labels are predicted rather than manually collected.

5 Conclusion

We proposed a statistical framework for lexical semantic change detection that models temporal consistency in usage-similarity score distributions. By evaluating whether the observed similarity structure is better explained by a single sense distribution or by distinct distributions across periods, our method provides a simple yet effective decision rule. Experiments on SemEval-2020 Task 1 and the WUGS datasets show that our approach achieves competitive or superior accuracy across languages and settings. These findings demonstrate that modeling usage-level similarity scores enables effective detection of semantic change. Together with recent advances in automatically predicting high-quality similarity labels, our framework offers a promising path toward fully automatic and interpretable semantic change detection.

References

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, 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.

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

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.

Paul Cook and Suzanne Stevenson. 2010. Automatically identifying changes in the semantic orientation of words. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

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

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908, 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.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *WWW 2015*, pages 625–635.

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.

Thomas P. Minka. 2000. Estimating a Dirichlet distribution. https://tminka.github.io/papers/dirichlet/.

Kevin P. Murphy. 2022. *Probabilistic Machine Learning: An Introduction*. MIT Press.

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.

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

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.

David Rother, Thomas 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*, pages 187–193, Barcelona (online). International Committee 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, 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, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association 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.

Dominik Schlechtweg, Shafqat Mumtaz Virk, Pauline Sander, Emma Sköldberg, Lukas Theuer Linke, Tuo Zhang, Nina Tahmasebi, Jonas Kuhn, and Sabine Schulte Im Walde. 2024. The DURel annotation tool: Human and computational measurement of semantic proximity, sense clusters and semantic change. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 137–149, St. Julians, Malta. Association for Computational Linguistics.

Zhaochen Su, Zecheng Tang, Xinyan Guan, Lijun Wu, Min Zhang, and Juntao Li. 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.

Elizabeth Closs Traugott and Richard B. Dasher. 2001. *Prior and current work on semantic change*, page 51–104. Cambridge Studies in Linguistics. Cambridge University Press.

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *WSDM 2018*, page 673–681.