ChiWUG: A Graph-based Evaluation Dataset for Chinese Lexical Semantic Change Detection

Jing Chen
The Hong Kong Polytechnic University
jing95.chen@connect.polyu.hk

Emmanuele Chersoni
The Hong Kong Polytechnic University
emmanuele.chersoni@polyu.edu.hk

Dominik Schlechtweg
University of Stuttgart
schlecdk@ims.uni-stuttgart.de

Jelena Prokic
Leiden University
j.prokic@hum.leidenuniv.nl

Chu-Ren Huang
The Hong Kong Polytechnic University
churen.huang@polyu.edu.hk

Abstract

Recent studies suggested that language models are efficient tools for measuring lexical semantic change. In our paper, we present the compilation of the first graph-based evaluation dataset for semantic change in the context of the Chinese language, covering the periods before and after the *Reform and Opening Up*.

Exploiting the existing framework DURel, we collect over 61,000 human semantic relatedness judgments for 40 targets. The inferred word usage graphs and semantic change scores provide a basis for visualization and evaluation of semantic change.

1 Introduction

Lexical semantic change detection, i.e. measuring meaning changes across different timespans, gained substantial popularity with the growing availability of historical corpora and language models (Hamilton et al., 2016; Tahmasebi et al., 2019; Montanelli and Periti, 2023; Kutuzov et al., 2018; Schlechtweg et al., 2020; Zamora-Reina et al., 2022), mostly for English and for other Indo-European languages.

The increasing number of published evaluation datasets further fostered the domain, enabling different models and hyperparameters to be quantitatively tested on the same benchmarks (Kutuzov et al., 2022; Schlechtweg et al., 2021; Aksenova et al., 2022; Chen et al., 2022; Zamora-Reina et al., 2022; Basile et al., 2019). These datasets are predominantly constructed within the framework of Diachronic Usage Relatedness (DURel), wherein changing scores are generated by calculating human ratings on semantic relatedness across a variety of usage pairs for targets (Schlechtweg et al., 2018; Rodina and Kutuzov, 2020; Chen et al., 2022). In the extended DURel framework, namely Diachronic Word Usage Graphs (DWUGs) (Schlechtweg et al., 2021, 2020), the usages could be further populated through *Word Usage Graphs* (WUGs) for visualization (McCarthy et al., 2016; Kutuzov et al., 2022).

To foster the development of lexical semantic change detection in Chinese, we constructed the first graph-based evaluation dataset, namely *ChiWUG*, following the DURel framework for the human judgments collection. Based on the collected 61k human judgments for 40 targets, we populated 40 WUGs to visualize usage changes preceding and following the context of the *Reform and Opening Up*, one of the most important milestones in the recent history of China. [1]

2 Related Work

Instead of categorizing words into *changed* and *unchanged* (Basile et al., 2020; Tang et al., 2013, 2016), the DURel framework adopted a graded view towards semantic change that words may exhibit varying degrees of semantic change. This is achieved by comparing the semantic relatedness targets in usage pairs on a scale of 1 to 4 (Schlechtweg et al., 2018), referring to semantic proximity from homonymy to identical usages. Specifically, usage pairs are assembled with contexts from periods of interest.

In the original DURel framework, three groups of usage pairs are assembled for a two-period setting, pairs consisting of two sentences from the same period and pairs having usages from each period (Schlechtweg et al., 2018). The extended

[1] The Reform and Opening Up period coincided with a series of policies implemented around 1978 to modernize the Chinese economy and engage with the global market.

Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change, pages 93–99

DWUGs allow us to categorize these usages into different groups by clustering, and groups proximately refer to different senses of a word. Through comparing the derived clusters from periods of interest, *DWUGs* allows us to easily measure the changes of sense distributions, i.e. loss and gain of senses and the turnover of usage dominance, which goes beyond the pure 'degree of changes' offered by the original DURel framework (Schlechtweg et al., 2021).

The DURel framework and its extension DWUGs have been applied to constructing evaluation datasets for a variety of languages, such as English, Swedish, German, and Latin released in the *SemEval 2020* (Schlechtweg et al., 2020), and later for Russian, Norwegian, Spanish, and Chinese (Rodina and Kutuzov, 2020; Kutuzov and Pivovarova, 2021; Kutuzov et al., 2022; Zamora-Reina et al., 2022; Chen et al., 2022). Since the nature of this paradigm is to measure usage differences between sentence pairs, it has also been extended to the construction of synchronic disambiguation datasets (Aksenova et al., 2022; Hätty et al., 2019) and to diatopic variation (i.e., usage differences across regional variations) (Baldissin et al., 2022).

3 Data

Building on the previous work by Chen et al. (2022), which collected human judgments for 20 targets following the DURel framework, we expand the data size and obtain the DWUGs to have a more comprehensive evaluation dataset for Chinese.[2]

3.1 Corpus

The corpus exploited in this study is derived from *People's Daily* [3], one of the most popular newspapers in China, which covers a wide range of topics. It is, to our knowledge, the largest continuous dataset with significant diachronic coverage that can be freely accessed. It covers the period from 1954 to 2003. All newspaper articles are in a Markdown format and are sorted into different temporal folders based on the release date.

More specifically, we take the year of the *Reform and Opening Up* as the borderline and divide all coverage into two subcorpora according to the releasing date information. One subcorpus contains all coverages from 1954 to 1978, and the other from 1979 to 2003. Table 1 summarizes word token/type information for the two sub-corpora. [4]

Period	Word Token	Word Type	TTR
1954 – 1978	1.27×10^8	46,743	0.368
1979 – 2003	1.66×10^8	58,376	0.351

Table 1: Statistics of two subcorpora. TTR = Type-Token ratio (Types/Tokens * 1000)

3.2 Target Words

To select targets, we first consulted Chinese linguistic studies on semantic change, with an emphasis on the period proceeding and following *Reform and Opening Up* (刁晏斌, 1995; 林伦伦, 2000; 于根元, 1992, 1994; 熊忠武, 1982; Tang et al., 2013, 2016; Tang, 2018). Considering the size and genre of our historical dataset, we only kept these candidates with validated senses recorded in the dictionaries. We do so by checking whether the mentioned emergent senses/usages were stabilized and absorbed into the standard Mandarin, relying on one of the most influential dictionaries in Modern Chinese (Department of Chinese Lexicography, 2019).

For example, '病毒'(bingdu, *virus*) developed a new sense roughly in the 1970s, relating to the computer virus, due to the introduction of the computer into the Chinese market (刁晏斌, 1995; Hamilton et al., 2016). However, 困难 'kun nan, *difficulty*' was recorded its usage as 'unattractive appearance' (刁晏斌, 1995), while such usage is neither much attested in the data nor recorded in dictionaries.

We further filtered those candidates with a normalized frequency of less than 1 in each period, specifically one from 1954 to 1978 (the EARLIER period) and the other from 1979 to 2003 (the LATER period).

Through such procedures, we identified a list of 20 changed words recorded in the linguistic literature as targets for constructing our evaluation dataset. Specifically, the list contains 11 verbs, 4 adjectives, and 5 nouns. We also selected an equal number of filler words as negative examples, only considering words of the same part of speech and comparable frequency in each period. Meanwhile, the same semantic field, with reference to

[2]Find the dataset at: https://zenodo.org/records/10023263.

[3]The *People's Daily Newspaper* Dataset: https://github.com/fangj/rmrb.

[4]Words averaging less than one occurrence in a one million tokens sample would be removed.

the dictionary 'Tongyici Cilin'(梅家驹, 1984), is also preferred if the first two criteria are met.

In sum, the evaluation dataset has 40 targets, including 20 changed words and 20 filler words. The changed words in this version have 9 out of 10 changed words in Chen et al. (2022) and the left one was filtered out due to frequency constraints. In general, the current *ChiWUG* dataset doubled the size of the target words.

3.3 Usage Pairs

To obtain semantic change scores, we first contextualized target words by providing actual usages in the historical dataset introduced in Section 3.1 and then asked native speakers to judge usage differences in the compared settings.

We first randomly sampled 40 sentences containing a target in each period from the dataset as sentence candidates and then removed those with insufficient contexts or/and having word segmentation errors after manual checking by the first author. In total, each target word has two groups of sampled sentences, containing 20 sentences from the EARLIER period and 20 from the LATER period [5]. Table 2 summarizes the general statistics regarding the sampled sentence pairs.

In theory, each sampled sentence would be automatically paired with each one of the other 39 sentences for comparison in the DWUG paradigm. Therefore, each target would have $(n(n-1))/2$ pairs, i.e. $(40 * 39)/2$, 780 pairs to be compared.

Targets	Sentences	Pairs	Avg Tokens per Sent.
40	1600	31,200	53.39

Table 2: Statistics of usage. *Avg Tokens per Sent.* refers to the average number of characters in sampled sentences.

4 Human Annotation

To collect human judgments, we recruited four native speakers of Mandarin as annotators. All the annotators are graduate students from the Faculty of Humanities, specializing in Chinese Linguistics. They were invited to experiment on the DURel platform after passing a tutorial specific to the Chinese lexical semantic change task [6]. Before the tutorial, we arranged a meeting for instructions.

After passing the tutorial, annotators were asked to indicate their intuitions on how semantically related a target was used in two displayed contexts in the 'official' annotation work. Targets would be highlighted, and options for judgments on a scale from identical (完全一致) to unrelated (不相关) are listed in the left bar, as shown in Figure 1. [7].

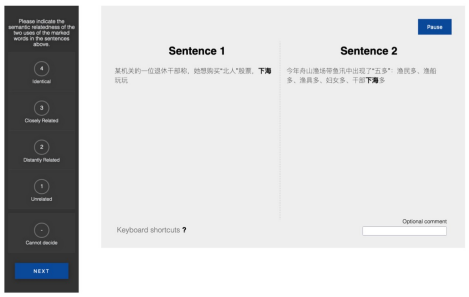

Figure 1: The annotation interface for Chinese

Besides assigning scores from 1 (unrelated meanings) to 4 (identical meanings) for semantic relatedness, they are also allowed to 'discard' usage pairs by giving the '0' score if the current pair is hard to understand due to the ambiguity of contexts or word segmentation errors. They are also encouraged to 'pause' the annotation process during the annotation after a period of annotation (around 30 minutes) to avoid excessively long sessions and keep their judgments as consistent as possible.

Due to the heavy load of annotation, the data was split in half, and each pair of annotators took one half consisting of 10 changed words and 10 fillers for annotation. We finally collected over 61,000 judgments from four annotators, after removing those judgments with a score of zero, that is, discarded pairs.

The weighted mean pairwise Spearman score for inter-rater agreements is 0.691, and the Krippendorff's alpha is 0.602, which are quite high if compared to other DURel datasets (Schlechtweg et al., 2021, 2020, 2018; Erk et al., 2013; Chen et al., 2022). For more statistics, see Table 3.

5 Graph Representations

Based on human judgments collected from the procedures described previously, we follow Schlechtweg et al. (2021, 2020) to aggregate the scores per usage pair as their median for populating

[5]The temporal information for all sentences is recorded in the meta-data, but would be invisible to annotators.

[6]The DURel interface: https://durel.ims.uni-stuttgart.de/. For the Chinese version of the guideline of this task: https://durel.ims.uni-stuttgart.de/guidelines?lang=ch

[7]More details: https://durel.ims.uni-stuttgart.de/guidelines?lang=zh

Periods	n	N/V/A	|U|	AN	JUD	AV	SPR	K
1954-2003	40	10/22/8	1,599	4	61k	2	.691	.602

Table 3: Statistics of target words in ChSemShift. n = the number of usages, $N/V/A$ = the number of nouns, verbs and adjectives, $|U|$ = the total number of usages. One usage pair was discarded during the annotation due to the context ambiguity. AN = the number of annotators, JUD = the number of judgments, AV = the average number of annotations per usage pair, SPR = weighted mean of pairwise Spearman score, K = Krippendorff's alpha.

WUGs, where usages with the same senses would be grouped together by performing the correlation clustering (Bansal et al., 2004). To populate WUGs with dense clusters, we took usage pairs with scores 3 and 4 as the same sense, while scores 1 and 2 were considered as different senses.[8]

Figure 2 and Figure 3 are inferred word usage graphs for 病毒 *bingdu* ('virus') and 下海 *xiahai* ('go into the sea' or 'to venture'), respectively. Nodes in the same color are clustered as the same sense, and subgraphs from the left to the right show the clusters/senses changes. In Figure 2, the right subgraph reveals the emergence of a new usage denoted by nodes in orange. By delving into the contexts associated with the orange-labeled usages [9], we discern that 病毒 (*bingdu*) acquired a fresh sense, namely 'computer virus', during the second period, diverging from its earlier associations with 'viral infections'.

Similarly, Figure 3 highlights the emergence of a new sense characterized by orange nodes, which are later confirmed as *'to venture'*, besides its original sense *'go into the sea'*. Furthermore, this emergent sense exhibits increased usage dominance in our samples, as evidenced by the greater presence of orange nodes in the LATER period.

6 Quantifying Changes: Metrics for Semantic Change

The DWUG paradigm obtains both graded change scores and binary change. We utilize two graded metrics: the Jensen-Shannon Distance on cluster frequency distributions (usually referred to as "graded change") as well as the COMPARE metric, calculated solely from edge weights. Binary change is instead based on the presence or absence of clusters across the two periods (Schlechtweg et al., 2018, 2020; Zamora-Reina et al., 2022).

[8]We use the WUG pipeline with default opt parameters to generate graphs, cluster them and compute statistics and change scores: https://github.com/Garrafao/WUGs.

[9]Clicking the nodes in the DURel platform would display the full context embedded in each node.

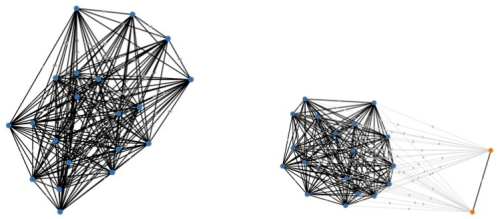

Figure 2: Word Usage Graphs of 病毒 *bingdu, 'virus'* in the EARLIER period *left* and the LATER period *right*). Colors label different clusters/senses, and nodes are different usages.

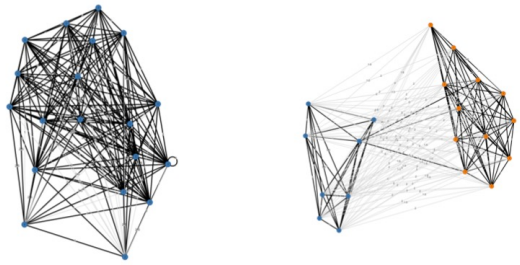

Figure 3: Word Usage Graphs of 下海 *xiahai, 'go into the sea' or 'to venture'* in the EARLIER period *left* and the LATER period *right*). Colors label different clusters/senses, and nodes are different usages.

COMPARE Metric The COMPARE metric C was proposed to directly compare the mean of weights where usages are from two different periods $W_{1,2}$, as shown in Eq. (1). A higher value yielded from the COMPARE metric indicates more stable words, while a lower value suggests a higher degree of meaning change (Schlechtweg et al., 2018; Schlechtweg, 2023).

$$C(W_{1,2}) = \frac{1}{|W_{1,2}|} \sum_{x \in W_{1,2}} x \qquad (1)$$

Jensen-Shannon Distance (Graded Change) After populating clusters, the frequency of sense distributions in two periods can be easily identified. To quantify the probability changes of sense distributions, the Jensen-Shannon Distance (JSD) is adopted to measure the change score between two normalized cluster frequency distributions (Schlechtweg, 2023), as shown in Eq. (2). The JSD is the symmetrized square root of the Kullback-Leibler Divergence (Lin, 1991). A higher JSD indi-

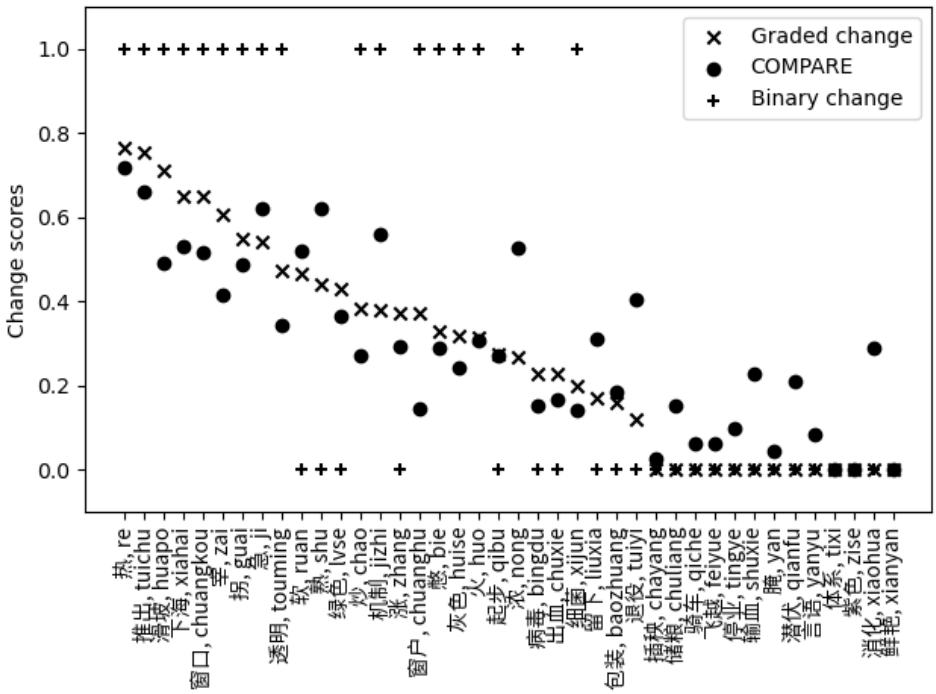

Figure 4: Change scores inferred on the WUGs resulting from our annotation. The COMPARE score was mapped with $f(x) = 1 - \frac{1}{3}(x-1)$ to fit the range of the other scores and to follow their direction (higher values mean more change).

cates a higher degree of usage change while a lower one suggests more stable usage across periods of interest.

$$JSD(P,Q) = \sqrt{\frac{KLD(P\|M) + KLD(Q\|M)}{2}} \quad (2)$$

where:

$$KLD(P\|Q) = \sum_{i}^{K} log_2(\frac{p_i}{q_i}), \quad M = \frac{(P+Q)}{2}$$

Binary Change The DWUG paradigm also enables us to detect binary change, defined as the gain or loss of clusters/senses. It is defined as:

$$B(w) = \begin{cases} 1 & \text{if for some } i, D_i \leq k \text{ and } E_i \geq n, \\ & \text{or vice versa.} \\ 0 & \text{otherwise} \end{cases}$$

where D_i and E_i respectively the frequency of sense i in the two periods, and k and n are lower frequency thresholds to control the handling of noise (Schlechtweg et al., 2020), which we set to $k = 1$ and $n = 3$.

Figure 4 demonstrates the change scores obtained on our data. Graded change and the COMPARE metric are strongly correlated (cf. Schlechtweg, 2023, pp. 63–64). Both scores in turn correlate with binary change. However, examples such as 软 *ruan* show that without binary change there can be considerable graded change. Similarly, words showing binary change can have varying degrees of graded change: 下海 *xiahai, 'go into the sea* or '*to venture'* demonstrate higher graded change than e.g. 病毒 'bingdu, *virus*' in Figure 4. Figure 2 and Figure 3 demonstrate their gaining of a new sense, respectively.

7 Conclusion

This study presents the first graph-based evaluation dataset for Chinese lexical semantic change constructed following the DWUG paradigm. It populates 40 word usage graphs based on more than 61k human judgments on contextual semantic relatedness between sentence pairs.

With its comparably high inter-rater agreement and dense clusters post-processed by clustering, we assume this high-quality evaluation dataset could be included in the shared evaluation datasets to foster Lexical semantic change detection in Chinese. Meanwhile, the inferred WUGs themselves are interesting for linguistic studies.

Limitations

We acknowledge that the periods we investigated were confined to a relatively short period of Chinese history, primarily spanning from the 1950s to the 2000s. Moreover, the analysis was concentrated on a specific regional source, utilizing a dataset derived from newspapers. While this scope is sufficient to unveil certain changes, it's imperative to acknowledge that the observed changes might merely represent a fraction of the broader evolutionary path. Changes identified within the current dataset could potentially be magnified or narrowed when explored within alternative data sources.

Acknowledgments

Dominik Schlechtweg has been funded by the project 'Towards Computational Lexical Semantic Change Detection' supported by the Swedish Research Council (20192022; contract 2018-01184) and by the research program 'Change is Key!' supported by Riksbankens Jubileumsfond (under reference number M21-0021).

References

Anna Aksenova, Ekaterina Gavrishina, Elisey Rykov, and Andrey Kutuzov. 2022. Rudsi: graph-based word sense induction dataset for russian. *arXiv preprint arXiv:2209.13750*.

Gioia Baldissin, Dominik Schlechtweg, and Sabine Schulte im Walde. 2022. DiaWUG: A Dataset for Diatopic Lexical Semantic Variation in Spanish. In *Proceedings of the 13th Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.

Nikhil Bansal, Avrim Blum, and Shuchi Chawla. 2004. Correlation clustering. *Machine Learning*, 56(1).

Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020. DIACR-Ita @ EVALITA2020: Overview of the EVALITA2020 Diachronic Lexical Semantics (DIACR-Ita) Task. In *Proceedings of EVALITA*.

Pierpaolo Basile, Giovanni Semeraro, and A. Caputo. 2019. Kronos-it: a dataset for the italian semantic change detection task. In *Italian Conference on Computational Linguistics*.

Jing Chen, Emmanuele Chersoni, and Chu-ren Huang. 2022. Lexicon of changes: Towards the evaluation of diachronic semantic shift in Chinese. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 113–118, Dublin, Ireland. Association for Computational Linguistics.

Chinese Academy of Social Science Department of Chinese Lexicography, Institute of Linguistics. 2019. *Contemporary Chinese Dictionary (Xiandai Hanyu Cidian)*, the 7th edition. Commercial Press, Peking.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2013. Measuring word meaning in context. *Computational Linguistics*, 39(3):511–554.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change.

Anna Hätty, Dominik Schlechtweg, and Sabine Schulte im Walde. 2019. SURel: A gold standard for incorporating meaning shifts into term extraction. In *Proceedings of the 8th Joint Conference on Lexical and Computational Semantics*, pages 1–8, Minneapolis, MN, USA.

Andrey Kutuzov and Lidia Pivovarova. 2021. Threepart Diachronic Semantic Change Dataset for Russian. In *Proceedings of the ACL International Workshop on Computational Approaches to Historical Language Change*.

Andrey Kutuzov, Samia Touileb, Petter Mhlum, Tita Ranveig Enstad, and Alexandra Wittemann. 2022. Nordiachange: Diachronic semantic change dataset for norwegian.

Andrey Kutuzov, Lilja vrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey.

Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37:145–151.

Diana McCarthy, Marianna Apidianaki, and Katrin Erk. 2016. Word sense clustering and clusterability. *Computational Linguistics*, 42(2):245–275.

Stefano Montanelli and Francesco Periti. 2023. A survey on contextualised semantic shift detection.

Julia Rodina and Andrey Kutuzov. 2020. RuSemShift: a dataset of historical lexical semantic change in Russian. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*. Association for Computational Linguistics.

Dominik Schlechtweg. 2023. *Human and Computational Measurement of Lexical Semantic Change*. Ph.D. thesis, University of Stuttgart, Stuttgart, Germany.

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 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. 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*, pages 169–174, New Orleans, Louisiana.

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.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2019. Survey of computational approaches to lexical semantic change.

Xuri Tang. 2018. A state-of-the-art of semantic change computation. *Natural Language Engineering*, 24(5):649–676.

Xuri Tang, Weiguang Qu, and Xiaohe Chen. 2013. Semantic change computation: A successive approach. In *Behavior and Social Computing*, pages 68–81, Cham. Springer International Publishing.

Xuri Tang, Weiguang Qu, and Xiaohe Chen. 2016. Semantic change computation: A successive approach. *World Wide Web*, 19.

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 International Workshop on Computational Approaches to Historical Language Change*, Dublin, Ireland. Association for Computational Linguistics.

于根元. 1992. *1991*汉语新词语. 北京语言学院出版社.

于根元. 1994. 现代汉语新词词典. 北京语言院出版社.

刁晏斌. 1995. 新时期大陆汉语的发展与变革. Hung Yeh Publishing, Taibei.

顾向欣 林伦伦, 朱永锴. 2000. 现代汉语新词语词典, *1978-2000*. 花城出版社.

梅家驹. 1984. 同林:. 商印; 上海.

熊忠武. 1982. 当代中国流行语词典. 吉林文史出版社.