The LSCD Benchmark: a Testbed for Diachronic Word Meaning Tasks

Dominik Schlechtweg[1], Sachin Yadav[1], Jonas Kuhn[1], Nikolay Arefyev[2]
University of Stuttgart[1], University of Oslo[2]
first.last@{ims.uni-stuttgart.de}

Abstract

Lexical Semantic Change Detection (LSCD) is a complex, lemma-level task, which is usually operationalized based on two subsequently applied usage-level tasks: First, Word-in-Context (WiC) labels are derived for pairs of usages. Then, these labels are represented in a graph on which Word Sense Induction (WSI) is applied to derive sense clusters. Finally, LSCD labels are derived by comparing sense clusters over time. This **modularity** is reflected in most LSCD datasets and models. It also leads to a large **heterogeneity** in modeling options and task definitions, which is exacerbated by a variety of dataset versions, preprocessing options and evaluation metrics. This heterogeneity makes it difficult to evaluate models under comparable conditions, to choose optimal model combinations or to reproduce results. Hence, we provide a benchmark repository standardizing LSCD evaluation. Through transparent implementation results become easily reproducible and by standardization different components can be freely combined. The repository reflects the task's modularity by allowing model evaluation for WiC, WSI and LSCD. This allows for careful evaluation of increasingly complex model components providing new ways of model optimization.

1 Introduction

Lexical Semantic Change Detection (LSCD) is a field of NLP that studies methods automating the analysis of changes in word meanings over time. In recent years, this field has seen much development in terms of models, datasets and tasks (de Sá et al., 2025; Periti and Montanelli, 2024), motivated by applications in historical linguistics or the digital humanities for theory testing (Hamilton et al., 2016; Karjus, 2025), or in lexicography for practical dictionary maintenance (Sköldberg et al., 2024). LSCD is a complex, lemma-level task, which is usually operationalized based on two subsequently applied usage-level tasks: First, Word-in-Context (WiC) labels are derived for pairs of usages. Then, these labels are represented in a graph on which Word Sense Induction (WSI) is applied to derive sense clusters. Finally, LSCD labels are derived by comparing sense clusters over time. This **modularity** is reflected in most LSCD datasets and models. It also leads to a large **heterogeneity** in modeling options and task definitions, which is exacerbated by a variety of dataset versions, preprocessing options and evaluation metrics. This heterogeneity makes it difficult to evaluate models under comparable conditions, to choose optimal model combinations or to reproduce results.

In order to handle this heterogeneity, we think that a shared testbed with a common evaluation setup is needed. Hence, we present a benchmark repository implementing evaluation procedures for models on most available LSCD datasets.[1] The benchmark exploits the modularity of the meta task LSCD by allowing for evaluation of the subtasks WiC and WSI on the same datasets. It can be assumed that performance on the subtasks directly determines performance on the meta task. We aim to stimulate transfer between the fields of WiC, WSI and LSCD by providing a repository allowing for evaluation on all these tasks with shared model components.

2 Related Work

A number of recently created LSCD datasets apply WiC and WSI in the annotation process (cf. Section 3) and thus allow for evaluation of WiC and WSI along with LSCD models (i.a. Kurtyigit et al., 2021; Schlechtweg et al., 2021; Kutuzov et al., 2022; Zamora-Reina et al., 2022; Chen et al., 2023).[2] There are also a number of datasets omit-

[1]Find the code at https://github.com/Garrafao/LSCDBenchmark.

[2]The bulk of these datasets is listed at: https://www.ims.uni-stuttgart.de/data/wugs.

*Proceedings of the 15th Joint Conference on Lexical and Computational Semantics (*SEM 2026)*, pages 148–163

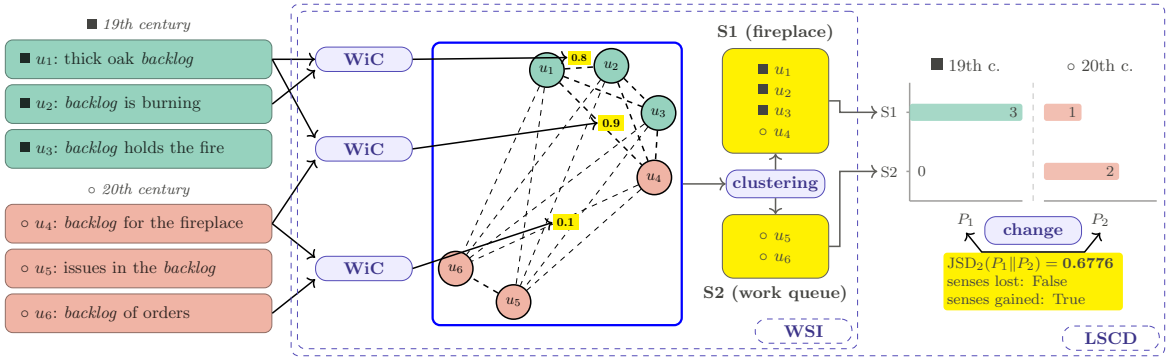

Figure 1: The straightforward approach to LSCD mimicking the annotation process of the popular LSCD datasets.

ting WSI, but allowing for WiC and LSCD evaluation (i.a. Schlechtweg et al., 2018; Rodina and Kutuzov, 2020; Kutuzov and Pivovarova, 2021), or datasets omitting the WiC allowing for WSI and LSCD evaluation (i.a. Basile et al., 2020; Cook et al., 2014), or datasets allowing purely for WiC evaluation (Loureiro et al., 2022).

So far, there is no comprehensive LSCD benchmark, implementing SOTA models on (human-annotated) high-quality evaluation data from multiple languages and multiple time periods. The leaderboards of several standalone shared tasks can be seen as small-scale benchmarks without common model implementation (Ahmad et al., 2020; Basile et al., 2020; Kutuzov and Pivovarova, 2021; Zamora-Reina et al., 2022; Fedorova et al., 2024) with the SemEval task being the most diverse with four languages (Schlechtweg et al., 2020). Schlechtweg et al. (2019) provide a comprehensive repository of type-based modeling approaches to LSCD with evaluation pipelines on multiple datasets. However, type-based models have more recently been outperformed by token-based contextualized embedding approaches (Kutuzov and Pivovarova, 2021; Zamora-Reina et al., 2022) and especially WiC-fine-tuned models (Cassotti et al., 2023; Homskiy and Arefyev, 2022). LLMs have shown competitive performance on some datasets (Zamora-Reina et al., 2025) while being less efficient limiting their applicability in large data scenarios. Periti and Tahmasebi (2024) perform a systematic comparison, but do not provide a flexible model implementation and do not evaluate on some of the most recent SOTA models or datasets.

3 Tasks

LSCD can be seen as the combination of (at least) three lexical semantic tasks (Schlechtweg, 2023): (i) measurement of semantic proximity between word usages, (ii) clustering of the usages based on their semantic proximity, and (iii) estimation of semantic change labels from the obtained clusterings. Task (i) and (ii) corresponds to the lexicographic process of deriving word senses (Kilgarriff, 2007), while task (iii) measures LSC based on the derived word senses. The tasks need to be solved sequentially, in the order given above, as each is dependent on the output of the previous task, e.g., word usages can only be clustered once their semantic proximity has been estimated.

The three tasks are reflected in the human (e.g. Schlechtweg et al., 2020, 2021; Kutuzov et al., 2022) as well as the computational process (e.g. Giulianelli et al., 2020; Montariol et al., 2021; Laicher et al., 2021; Homskiy and Arefyev, 2022) of measuring lexical semantic change.[3] The first task is known as a standalone task under the name of "Word-in-Context" (WiC, Pilehvar and Camacho-Collados, 2019) while the second task is known as the task of "Word Sense Induction" (WSI, Schütze, 1998). A number of recently created LSCD datasets reflect all of these tasks and thus allow for evaluation of WiC and WSI along with LSCD models (i.a. Schlechtweg et al., 2021; Kurtyigit et al., 2021; Kutuzov et al., 2022; Zamora-Reina et al., 2022; Chen et al., 2023).

Below, we provide detailed descriptions of each task. Figure 1 sketches the most straightforward and modular approach to LSCD which follows the annotation process of the popular LSCD datasets and explicitly solves each task sequentially: Usages of the target word *backlog* from two time periods are combined into pairs and for each pair a WiC label is predicted. (In this case it is a continuous score, but it could be binary or ordinal.) These are

[3] However, it is not always obvious as annotation and modeling procedures often try to simplify or skip steps of this process.

then represented in a graph and clustered. From the clusters, sense probability distributions for the two time periods are derived, which are then compared to measure change, e.g. by the JSD-Shannon Distance (Lin, 1991) for Graded Change or sense loss and gain for Binary Change. Our benchmark also provides implementations of other approaches to LSCD that either skip some of the tasks or solve them differently from the dataset annotation process.

3.1 Word-in-Context

WiC asks to determine if two words occurring in two text fragments have the same or different meanings. Usually two usages of the same word probably in different grammatical forms are given. For example, consider any pair of usages of the word *backlog* in Figure 1. The WiC task is often framed as a binary classification task. For instance, the WiC (Pilehvar and Camacho-Collados, 2019) and MCL-WiC (Martelli et al., 2021) datasets contain binary labels and employ accuracy as the main evaluation metric. Alternatively, USim (Erk et al., 2013), SCWS (Huang et al., 2012) and CoSimLex (Armendariz et al., 2020) were labeled with non-binary semantic proximity scores and promote a graded formulation of the task. In this formulation, a WiC model shall produce scores that are similar to the human scores, or at least rank the pairs of usages similarly. Spearman's and Pearson's correlation coefficients are employed as evaluation metrics in this case (Spearman, 1904). More recently, Schlechtweg et al. (2025) provided an ordinal formulation of the graded WiC task, evaluating with Krippendorff's α (Krippendorff, 2018).

Following the DURel annotation framework for LSCD (Schlechtweg et al., 2024b), human annotators solved the graded WiC task, i.e., they annotated the semantic proximity of two usages of the same word on a scale. This provides data for evaluation of WiC models that can serve as a part of LSCD models. In diachronic LSCD datasets, there are pairs of word usages extracted from two documents belonging to distant time periods making usages in these pairs very different orthographically, grammatically, and thematically, even when the target word has the same meaning. This might be challenging for models trained on traditional WiC datasets, which often contain examples from the same time period. Our benchmark helps to analyze how sensitive WiC models are to this shift in time period by comparing their performance on pairs of usages extracted from the old, the new or both corpora.

3.2 Word Sense Induction

WSI asks to infer which senses a given target word has based only on its usages in an unlabeled corpus. It is usually framed as a clustering task where a model shall cluster a given set of usages of the same target word probably in different grammatical forms into clusters corresponding to the senses of this word. One way to solve the task is through clustering a weighted graph populated with WiC predictions, cf. the middle part of Figure 1. Unlike the more popular Word Sense Disambiguation task, in WSI no sense inventory is given to the model and the number of senses of the target word is not known as well. The most widespread formulation of WSI assumes that each usage has one and only one sense, thus, requires hard clustering, i.e. assigning each usage to a single cluster (i.a. SemEval 2010 Task 14, Manandhar and Klapaftis, 2009).

3.3 Lexical Semantic Change Detection

LSCD is a general name for several tasks dealing with analysis of different properties of a word related to changes in its meaning over time. Usually, in these tasks a list of target words are given and two time periods are specified, an old and a new one. Each time period is represented by an unlabeled corpus or a pre-selected set of usages.

The Binary Change task (Schlechtweg et al., 2020; Zamora-Reina et al., 2022) asks if the set of senses of a given word is the same for two time periods, i.e., whether any senses were lost or gained, or not. It assumes that word meaning in a particular time period can be described as a set of discrete and mutually exclusive senses observed in the corresponding corpus. This task can be viewed as a task of binary classification of words. The Graded Change (or JSD) task (Schlechtweg et al., 2020; Zamora-Reina et al., 2022) has the same assumptions about word meaning, but instead of binary classification it requires ranking a given list of words according to changes in their sense frequency distributions. The rank of a word is determined by the Jensen–Shannon Distance between two probability distributions over word senses $P(sense|w, t_{old})$ and $P(sense|w, t_{new})$, one for the older time period and another for the newer one (Schlechtweg et al., 2020; Zamora-Reina et al., 2022). The Binary and Graded Change tasks are illustrated in the right part of Figure 1. A com-

Data set	LGS	n	N/V/A	|U|	AN	JUD	Task	t_1	t_2	Reference	Version
DWUG	DE	50	32/14/2	178	8	63k	WiC, WSI, LSCD (B,G,C)	1800–1899	1946–1990	Schlechtweg et al. (2021)	3.0.0
DWUG Res.	DE	15	10/4/1	50	3	10k	WiC, WSI, LSCD (B,G,C)	1800–1899	1946–1990	Schlechtweg et al. (2024a)	1.0.0
DWUG	EN	46	40/6/0	191	13	29k	WiC, WSI, LSCD (B,G,C)	1810–1860	1960–2010	Schlechtweg et al. (2021)	3.0.0
DWUG Res.	EN	15	14/1/0	50	3	7K	WiC, WSI, LSCD (B,G,C)	1810–1860	1960–2010	Schlechtweg et al. (2024a)	1.0.0
DWUG	SV	44	32/5/7	171	13	20k	WiC, WSI, LSCD (B,G,C)	1790–1830	1895–1903	Schlechtweg et al. (2021)	3.0.0
DWUG Res.	SV	15	10/3/2	50	6	16k	WiC, WSI, LSCD (B,G,C)	1790–1830	1895–1903	Schlechtweg et al. (2024a)	1.0.0
DWUG	ES	100	51/24/25	40	12	62k	WiC, WSI, LSCD (B,G,C)	1810–1906	1994–2020	Zamora-Reina et al. (2022)	4.0.2
DiscoWUG	DE	75	39/16/20	49	8	24k	WiC, WSI, LSCD (B,G,C)	1800–1899	1946–1990	Kurtyigit et al. (2021)	2.0.0
RefWUG	DE	22	15/1/6	19	5	4k	WiC, WSI, LSCD (B,G,C)	1750–1800	1850–1900	Schlechtweg (2023)	1.1.0
NorDiaChange1	NO	40	40/0/0	21	3	14k	WiC, WSI, LSCD (B,G,C)	1929–1965	1970–2013	Kutuzov et al. (2022)	1.0.0
NorDiaChange2	NO	40	40/0/0	21	3	15k	WiC, WSI, LSCD (B,G,C)	1980–1990	2012–2019	Kutuzov et al. (2022)	1.0.0
ChiWUG	ZH	40	10/22/8	40	4	61k	WiC, WSI, LSCD (B,G,C)	954-1978	1979-2003	Chen et al. (2023)	1.0.0
DWUG	IT	26	17/3/6	86	5	5k	WiC, WSI, LSCD (B,G,C)	1948-1970	1990-2014	Cassotti et al. (2024a)	3.0.0
DURel	DE	22	15/1/6	104	5	6k	WiC, LSCD (C)	1750–1800	1850–1900	Schlechtweg et al. (2018)	3.0.0
SURel	DE	22	19/3/0	104	4	5k	WiC, LSCD (C)	general	domain	Hätty et al. (2019)	3.0.0
RuSemShift1	RU	71	65/6/0	119	5	21k	WiC, LSCD (C)	1682–1916	1918–1990	Rodina and Kutuzov (2020)	2.0.0
RuSemShift2	RU	69	57/12/0	105	5	18k	WiC, LSCD (C)	1918–1990	1991–2016	Rodina and Kutuzov (2020)	2.0.0
RuShiftEval1	RU	111	111/0/0	60	3	10k	WiC, LSCD (C)	1682–1916	1918–1990	Kutuzov and Pivovarova (2021)	2.0.0
RuShiftEval2	RU	111	111/0/0	60	3	10k	WiC, LSCD (C)	1918–1990	1991–2016	Kutuzov and Pivovarova (2021)	2.0.0
RuShiftEval3	RU	111	111/0/0	60	3	10k	WiC, LSCD (C)	1682–1916	1991–2016	Kutuzov and Pivovarova (2021)	2.0.0

Table 1: Overview datasets. LGS = language, n = no. of target words, N/V/A = no. of nouns/verbs/adjectives, $|U|$ = avg. no. usages per word, AN = no. of annotators, JUD = total no. of judged usage pairs, Task = possible evaluation tasks, t_1, t_2 = time period 1/2, Reference = data set reference paper, Version = version used for experiments.

mon way to derive sense probability distributions is through WiC-based WSI, middle part of Figure 1.

Finally, there is also the COMPARE task (Schlechtweg et al., 2018; Kutuzov and Pivovarova, 2021; Zamora-Reina et al., 2022), which requires ranking a given set of words according to the average proximity (WiC score) between old and new usages of each word, i.e., the average cross-period proximity. This task avoids clustering and can be seen as a simple approximation of the Graded Change task.

4 Datasets

Table 1 shows all datasets currently integrated into the benchmark. All datasets have in common that they are based on human WiC judgments of word usage pairs on the ordinal DURel scale from 1 to 4, where 1 means semantically unrelated and 4 means identical (Schlechtweg et al., 2018). They also share the use of diachronic data.[4]

The datasets then fall into two main categories: (i) Datasets representing annotated judgments in a sparsely connected graph (Word Usage Graph, find an example in Figure 8, in Appendix C), clustering these with a variation of correlation clustering (Bansal et al., 2004; Schlechtweg et al., 2021), and deriving LSC labels by comparing the two time-specific sense frequency distributions (Schlechtweg et al., 2020), as illustrated in Figure 1. These datasets, displayed in the upper part of Table 1, apply the full lexicographic process and thus allow for full evaluation on all tasks mentioned in Section 3. (ii) Datasets skipping the clustering step and hence only allowing for evaluation on the WiC and COMPARE tasks. These are shown in the lower part of Table 1. Apart from the difference in tasks they support, the datasets have strongly varying properties in terms of language, number of target words, PoS distribution, number of usages per target word and number of human judgments. We integrate provided preprocessed usages (tokenization, lemmatization, normalization, PoS) as preprocessing options together with general strategies such as target word substitution by the lemma. There is a range of further data which could be integrated into the benchmark with minor adjustments such as WiC or Raw-C (Pilehvar and Camacho-Collados, 2019; Trott and Bergen, 2021).

5 Evaluation Procedures

Figure 2 shows the structure of our benchmark. It summarizes token-based approaches to LSCD and shows how their components and whole pipelines can be evaluated using our benchmark. The benchmark reflects the main LSCD approach illustrated in Figure 1, but also allows for simpler modeling, e.g. by skipping the clustering step.

The central part shows the WSI-based approach to LSCD. It relies on WSI methods which cluster word usages based either on their contextualized embeddings or pairwise similarities between them calculated by a WiC model. If a WiC model is involved, we can evaluate it separately on all datasets containing human-labeled pairs of word usages by feeding these pairs and comparing the model predictions with the human labels. Spearman's and Pearson's correlation coefficients that compare rankings or scores predicted by humans

[4]With the exception of SURel comparing usages from different domains rather than time periods (Hätty et al., 2019).

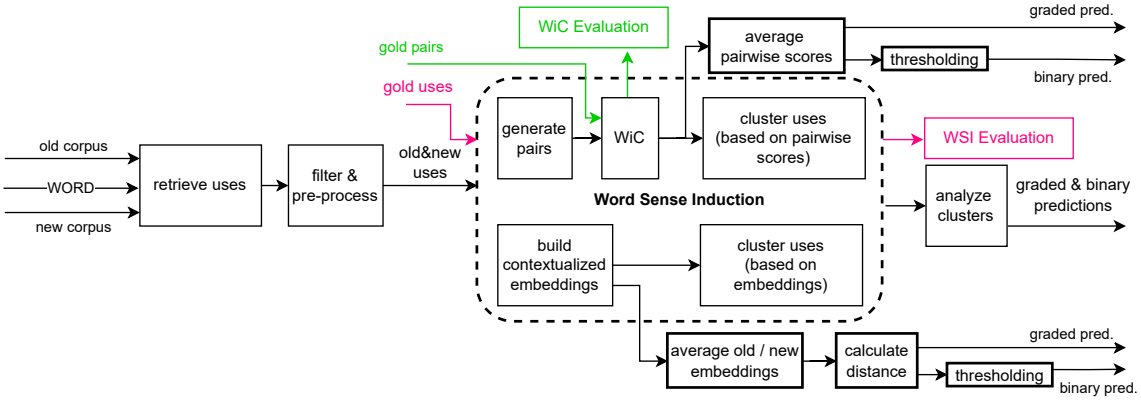

Figure 2: Token-based LSCD pipelines and their evaluation.

and the WiC model are employed as metrics for the WiC task. A WSI method can also be evaluated as a whole by running it on a set of gold usages of each word, i.e. usages that have sense labels obtained either directly from human annotators, or by clustering Word Usage Graphs. Clusterings obtained with the WSI method are compared against sense labels using the Adjusted Rand Index (Hubert and Arabie, 1985) as a the main metric. Finally, we can evaluate the whole LSCD pipeline using the standard LSCD metrics, i.e. F1-score for the binary classification tasks or Spearman's correlation with the gold word ranking for the JSD and COMPARE tasks.

We introduce standard splits for each dataset on the lemma level, i.e., certain target words are assigned to train/dev/test. We further provide possibility to evaluate on the previously introduced standard split from the CoMeDi shared task (Schlechtweg et al., 2025). Find an overview of the number of target words in this split in Appendix A.

6 Models

The most straightforward modeling approach for LSCD models is to follow the 3-level annotation approach of WUGs described in Section 3 and displayed in Figure 1. Given a target word, a basic model retrieves uses of this word from an *old* and a *new* corpus, then clusters them in order to infer word senses, and finally analyzes the obtained clusters to make conclusions about changes in word senses between two time periods. This approach is appealing because if word senses are inferred correctly, then an exact description of how word senses changed, as well as the exact predictions for all LSCD tasks are easy to obtain. Also the procedure basically automates the annotation procedure of various LSCD datasets, which is a reasonable way of getting predictions that correspond well to the ground truth. The benchmark implementation as depicted in Figure 2 thus follows this basic structure. The inputs and the outputs of each component are specified in round brackets, while the hyperparameters are underlined in the corresponding descriptions. The implemented components on each of these levels will be described below. It is important to note that current SOTA models for Graded Change (Giulianelli et al., 2020; Arefyev et al., 2021a; Cassotti et al., 2023) skip the clustering step and model Graded Change directly from WiC predictions of proximity between word usages or their underlying vectors by aggregating these time-wise (see below).

Retrieve uses (word, corpus → uses). Given a corpus and a target word, this component retrieves all uses of the target word in all of its grammatical forms from the given corpus. Optionally, N uses may be sampled if there are more than that in order to reduce the following computations.

For intrinsic evaluation, instead of retrieving uses, golden uses may be taken, i.e., the uses which were shown to the annotators during dataset construction. This eliminates possible disagreements due to some rare senses of the target word sampled during the dataset annotation procedure but not sampled by the model during evaluation, or vice versa.

Build contextualized embeddings (uses → embeddings). We employ BERT or XLM-R masked language *model* to obtain the contextualized representations of the given target word in the given text fragment (Devlin et al., 2019; Conneau et al., 2020).[5] The simplest and most popular option is

[5]In principle, the benchmark supports all model checkpoints from huggingface.

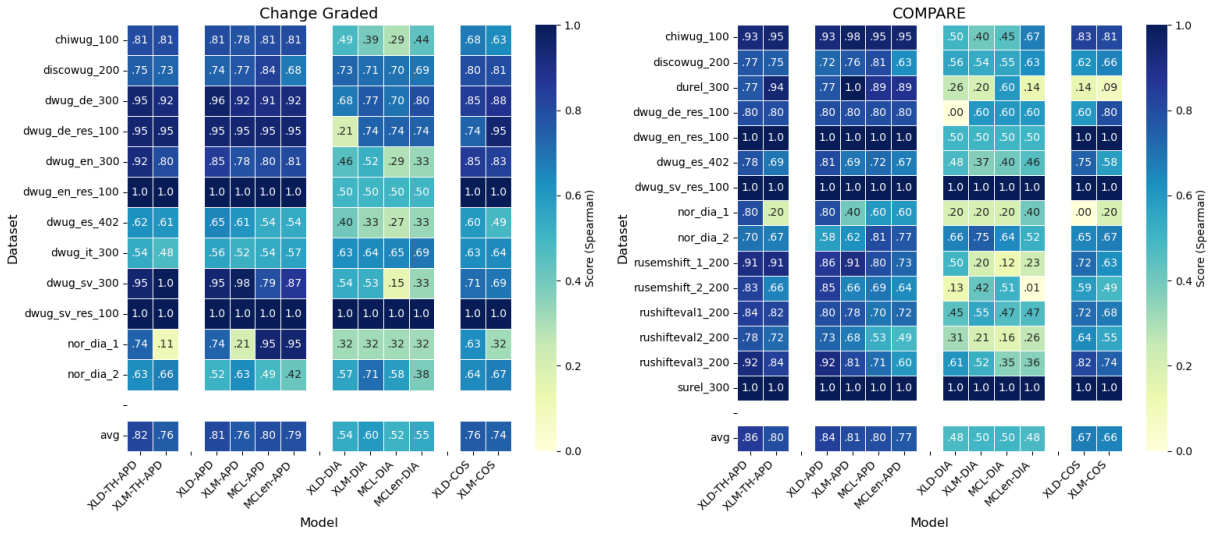

Figure 3: Result overview on Graded Change (left) and COMPARE score (right). XLD = XL-DURel, XLM = XL-LEXEME, MCL/MCLen = DeepMistake checkpoints, TH = thresholded, APD = Average Pairwise Distance, DIA = DiaSense, COS = Cosine distance between average embeddings.

taking the outputs of the last Transformer layer (before the MLM head) on the positions of the target word and average those outputs (mean pooling). We can apply other *subword poolings*, max pooling or first pooling (take the outputs from the first subword). A more general implementation combines the outputs from several *layers*. Again, the simplest option is just averaging them.[6] However, the experiments with BERT without fine-tuning presented in Devlin et al. (2019) for the NER task suggest that it may be better to concatenate the outputs of the last four layers instead of averaging them. Thus, the *layer aggregation function* is a hyperparameter and we select its value among averaging and concatenation. Additionally, the benchmark supports XL-LEXEME, a WiC-fine-tuned version of XLM-R (Cassotti et al., 2023) and the similar XL-DURel which is instead fine-tuned for ordinal WiC (Yadav and Schlechtweg, 2025).

Generate pairs (uses → pairs of uses). Pairs of uses of the specified *type* are generated. For the COMPARE type, each pair contains a use from the old corpus and a use from the new corpus. This type is ideal for the COMPARE task in which one needs to estimate the average proximity between old and new usages. If type is ALL, then all possible pairs are generated. This is useful to provide more information for the following clustering step.

WiC (pairs of uses → pairwise scores). We consider two different approaches to the WiC task. The first approach builds the contextualized embeddings for all uses with the aforementioned component, it inherits all the corresponding *hyperparameters*. Then, it employs one of the *distance functions* to calculate distances between the contextualized embeddings of two uses in each pair. The euclidean, manhattan and cosine distances are currently supported.

The second approach employs a binary classifier implemented as a neural network that jointly processes a pair of word usages and returns the probability that the meaning is the same. We treat this probability as the proximity between word occurrences. We experimented with the DeepMistake system[7] which had achieved the 2nd best result in the RuShiftEval and LSCDiscovery shared tasks (Arefyev et al., 2021a; Homskiy and Arefyev, 2022), and also with the method from (Yadav and Schlechtweg, 2025; Schlechtweg et al., 2025) which has shown good performance on the recent ordinal WiC shared task data.

We further provide the possibility of discretizing graded WiC predictions at specified thresholds. By default, the thresholds from Yadav and Schlechtweg (2025) are used.

[6] Check Vulić et al. (2020) for a detailed study on averaging layers for lexical tasks.

[7] DeepMistake is a family of WiC models. It includes models fine-tuned on the data from the LSCD shared tasks in Russian or Spanish, and more general WiC models trained on general-purpose multilingual WiC datasets which we employ in our study. Specifically, we selected the MCL and MCLen models trained on the whole or the English part of the MCL-WiC dataset respectively. We further compare them to the language-specific models in Appendix B.

Clustering (pairwise scores → clusters or contextualized embeddings → clusters). The goal of this step is to discover all senses of the target word occurring in the old, or the new corpus, or both of them, i.e., solve the WSI task for all uses retrieved from both corpora. Clusters are inferred simultaneously on all usages from old and new corpus. The clustering algorithms that can accept a matrix of similarities or distances between objects instead of object vectors can be applied to both types of inputs to the clustering component (pairwise scores, embeddings). However, some clustering algorithms require raw object vectors and can be applied to the contextualized embeddings, but not to the pairwise scores. For instance, K-Means calculates cluster centroids and needs object vectors to do that. The benchmark currently supports correlation clustering (Bansal et al., 2004) operating on pairwise scores. We chose this algorithm as most gold annotations were clustered with this approach (e.g. Schlechtweg et al., 2020).

Cluster measures (clusters → LSCD predictions). To solve the Binary Change task, after WSI we search for those clusters that contain only new or only old examples. Those clusters are viewed as novel or lost senses correspondingly, and the binary labels are predicted accordingly. Alternatively, we search for the clusters with at least M new examples and at most K old examples or vice versa in order to align with the annotation procedure of some LSCD datasets. To solve the JSD task, for old and new uses separately we estimate the probability distribution over senses of the target word as the proportions of its uses present in each cluster. Then the JSD between the two probability distributions is calculated.

Aggregate measures (pairwise scores → LSCD predictions or contextualized embeddings → LSCD predictions). These skip the clustering step by aggregating WiC predictions directly. The most successful of these is *Average Pairwise Distance (APD)*, which simply averages the distances or similarities between word usages from different time periods (COMPARE pairs) returned by some WiC model (Kutuzov and Giulianelli, 2020). This follows the calculation of the gold COMPARE scores, which is the average of human pairwise annotations. APD is sensitive to the polysemy or variation of the target word, which can lead to wrong predictions. For this, measures normalizing APD by a polysemy term were proposed, such as *DiaSense*

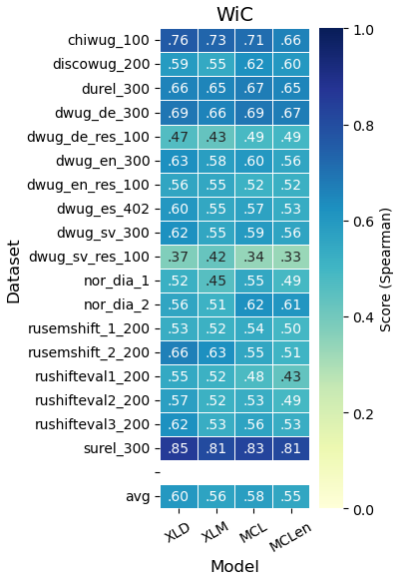

Figure 4: Result overview on WiC.

(Beck, 2020). *COS*, in contrast, averages the contextualized embeddings of all old and all new uses separately producing two aggregated embeddings of the target word for each of of the two time periods. Then, it calculates the distance between those two embeddings. It was previously introduced under the names of PRT (Kutuzov and Giulianelli, 2020) and COS (Laicher et al., 2021), cf. also Martinc et al. (2020). This distance is used as the prediction for both graded tasks (Graded Change and COMPARE).

7 Experiments

We now use the benchmark to perform a number of experiments. We focus on Graded Change and COMPARE detection as these are the most widely approached LSCD tasks (Kutuzov and Pivovarova, 2021; Periti and Tahmasebi, 2024). Note that not all datasets provide both evaluation scores (see Table 1). All experiments are performed on the CoMeDi test split described in Section 1. We cannot test all models and configurations, thus we focus on SOTA models using aggregate change measures without clustering and design experiments to answer open research questions.

Which WiC model and which aggregate measure gives SOTA performance? According to recent studies (Zamora-Reina et al., 2022; Periti and Tahmasebi, 2024; Zamora-Reina et al., 2025), DeepMistake and XL-LEXEME combined with APD compete for the SOTA on Graded Change and COMPARE detection. Previous studies have not directly compared these models though, or

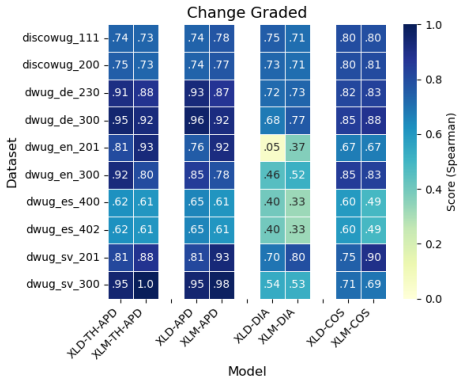

Figure 5: Result overview on dataset versions for bi-encoder models.

only on very limited data (Zamora-Reina et al., 2025). Hence, we perform a direct comparison guaranteeing a fair evaluation setup. We also include the recently published XL-DURel (Yadav and Schlechtweg, 2025), which has been shown to improve upon both models on ordinal WiC. In addition to the canonical datasets, we include a number of recently published datasets which have never been used for model evaluation: DWUG DE/EN/SV V3.0.0, DWUG DE/EN/SV resampled and DWUG IT. This is the most thorough model comparison done so far in the field of LSCD.

Figure 3 shows the results across the three SOTA WiC models, each combined with the common aggregate change measures APD, COS and DiaSense (see Section 6). Both, for Graded Change and COMPARE, clearly APD dominates which confirms previous results. However, there are slight advantages for the recently published XL-DURel model suggesting a new SOTA.

Does WiC prediction discretization improve results? All above-reported measures are based on aggregation of graded WiC predictions (either cosine similarity or same-sense probability). However, all WUG datasets are annotated on an ordinal (discrete) scale (see Section 4). Exactly predicting these ordinal values can be done by thresholding the graded predictions (Choppa et al., 2025). We hypothesize that discretizing WiC predictions to resemble human annotations helps for LSCD as ordinal judgments were used for ground-truth construction. Hence, in Figure 3 we report two models with APD and thresholding (XLD/XLM-TH-APD). Threshold parameters were taken from Yadav and Schlechtweg (2025). As we see, for XL-DURel, thresholding slightly helps, further pushing the SOTA, while for XL-LEXEME it has no effect or slightly hurts.

Which model gives SOTA on diachronic WiC? Does WiC determine LSCD? A recent shared task has compared DeepMistake and XL-LEXEME for ordinal WiC showing a slight advantage for DeepMistake (Schlechtweg et al., 2025). XL-DURel has further improved upon both models. Performance on the WiC task is a strongly influential factor for LSCD (Arefyev et al., 2021b). We compare model performance on these two levels to gain an understanding how strongly WiC determines LSCD performance. Figure 4 compares WiC models on the ordinal WiC task. The MCL checkpoint of DeepMistake does dominate XL-LEXEME confirming previous results, but both are outperformed by XL-DURel. Now compare this to the APD columns in Figure 3. Overall, WiC performance determines LSCD performance. Consider e.g. the dominance of XL-DURel on DWUG EN with Graded Change. However, there are notable exceptions such as DWUG SV where XL-DURel dominates on WiC but not LSCD. This shows that purely improving WiC ranking does not guarantee better LSCD performance. Possibly, the score distribution plays an important role when averaging values for aggregate measures.

Are model performances reproducible with more reliable data? What is the performance development on incrementally annotated datasets? Many datasets have been annotated in incremental rounds of annotation (Schlechtweg et al., 2021, 2024a). Later rounds are supposed to yield higher data quality as graphs are more richly annotated. Schlechtweg et al. (2024a) suggest that previous model comparisons done on older dataset versions/less rounds should be repeated with the more reliable data (last round). We are the first to investigate model performance with the latest dataset versions. This will also allow us to investigate the impact of annotation rounds on performance. Figure 5 shows the performance of a selection of models on consecutive versions. For the top models, we can see that performance tends to increase with later versions, which is expected as quality should increase. However, there are cases where model performance strongly drops for newer versions, such as XLM-APD on DWUG EN. Further, the relative performance of models can strongly change depending on version, e.g. XLM-APD outperforms XLD-APD on DWUG EN for an older version while it is the other way around for the newer, more reliable version. This shows the risks

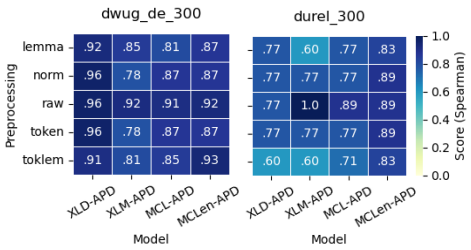

Figure 6: Result selection on preprocessing. lemma = lemmatization, norm = normalization, raw = no preprocessing, token = tokenization, toklem = tokenization with target word substituted by lemma.

of using unreliable ground-truth data and illustrates the need for the creation of benchmarks such as ours and continuous model reevaluation on additional and improved datasets.

What is the impact of spelling variation on performance? Laicher (2020) show that historical spelling variations can have a strong influence on BERT-based model performance. However, it is not clear how much this influences SOTA models. Hence, we test the influence of spelling normalization and lemmatization on two German datasets containing historical spelling variants in Figure 6. We can see that applying no preprocessing (raw) gives top performance for all models, with one exception (MCLen with toklem). This suggests that current base embedders are quite robust and do not need additional preprocessing.

Relation to previous work. We did not evaluate models on the full set of target words from each dataset, but only on the CoMeDi test split, as described in Section 5. This enabled us to test models such as XL-DURel, which was fine-tuned on the use-pair-level data of many LSCD datasets making it impossible to be tested on the full data. This also means that most previous results are not directly comparable to ours as they were obtained on the full (or almost full) datasets. On the WiC-level, the results from Yadav and Schlechtweg (2025) are weakly comparable as they were obtained on the same split, but with a stronger cleaning strategy and data concatenated per language. This leads to lower top results in our evaluation on German data (except for SURel), English and Spanish, higher results for Chinese and Norwegian, and mixed results for Swedish and Russian. Similarly, on the LSCD-level, the results from Periti and Tahmasebi (2024) are weakly comparable as they were obtained on the full datasets including the split we evaluated on. Our top results are higher for Chinese, German, English, Swedish, and similar for Russian, Spanish and Swedish. In the future, it will be essential (and also very simple) to compare models on the full data to make the results more comparable to previous work.

Results on concatenated data. We experimented with evaluating models on different concatenations of datasets and noticed that some model performance differences vanish in this setup. In a further analysis of model predictions, we realized that at least part of the reason for this effect are prediction outliers in individual datasets which deteriorate model performance much more in larger concatenations of data. While concatenating datasets is a good way to overcome data sparsity, we are unsure about the validity of this approach as it assumes that datasets were annotated consistently and that dataset-specific properties such as the size of the annotated graphs do not introduce biases that would make change scores incomparable. We leave more research in this direction to future work.

8 Conclusion

In this work, we have presented a new benchmark for evaluation of token-based LSCD models. The procedures are implemented that can evaluate both whole LSCD solutions and their separate components that solve the WiC and WSI subtasks. A variety of LSCD datasets are integrated in the benchmark allowing thorough evaluation on various languages and diverse historical epochs. We used the benchmark to perform a number of experiments with recent models setting a new SOTA in LSCD and providing a better general understanding on LSCD model evaluation and improvement.

Our main findings can be summarized as follows: (i) Recent WiC models optimized for ranking use pairs on an ordinal scale perform competitively or better than WiC models optimized on binary WiC data. (ii) Discretizing graded WiC predictions to ordinal, human-like values helps for ordinal WiC models. (iii) Overall, WiC performance determines LSCD performance. (iv) Using less reliable, older versions of datasets can lead to false conclusions on optimal model choice. (v) Current base embedder models are quite robust against spelling variation. Overall, our results suggest a kind of trivial, but rarely acknowledged trend in LSCD: More closely modelling the human annotation process leads to better performance.

Limitations

In our evaluation, we did not evaluate any cluster-based models although these have a large potential for high performance (Zamora-Reina et al., 2022; Schlechtweg et al., 2024c). However, the current SOTA does not build on clustering. Hence, we focused our evaluation to give a more concise overview and leave the comparison to cluster-based models to future work. We have also used seemingly non-optimal DeepMistake model training checkpoints, likely underestimating its performance. Another problem is the small size of the CoMeDi test split, potentially leading to strong performance variations through sampling error.

In the future, we want to use the benchmark to compare clustering models including a comparison on full datasets. We also want to add more datasets such as the Slovenian or the Japanese one (Pranjić et al., 2025; Ling et al., 2023; Baldissin et al., 2022), provided they are publicly available in the correct format. This will have the advantage that we can test all models on the full datasets as they have not been part of the training data. Further, we want to add more recently developed semantic change metrics (Pranjić et al., 2025; Goworek and Dubossarsky, 2026) and base embedders optimized for ordinal WiC prediction (Mujko, 2026). On the long term, we also see the possibility of integrating the benchmark with more fine-grained LSCD datasets providing semantic change types (Cassotti et al., 2024b; Whaley, 2025) or a definition generation component (Fedorova et al., 2024).

While our aim is to standardize LSCD model evaluation, we do not aim to delegitimize other evaluation approaches or variations of the one used here. In our benchmark, we try to capture as much heterogeneity as we can within the current main LSCD paradigm, but approaches focusing on types of change (Cassotti et al., 2024b; Whaley, 2025) or detecting changes against a dictionary (Fedorova et al., 2024; Blessing et al., 2026) are equally valid and should be pursued further, in our opinion.

Acknowledgments

Dominik Schlechtweg and Sachin Yadav have been funded by the research program 'Change is Key!' supported by Riksbankens Jubileumsfond (under reference number M21-0021). Nikolay Arefyev has received funding from the European Union's Horizon Europe research and innovation program under Grant agreement No 101070350 (HPLT). Thanks to Andres Cabero, Kuan-Yu Lin and Arshan Seyed Dalili for contributing code to the repository. Thanks to Shafqat Mumtaz Virk for contributing to an earlier version of this paper.

References

Adnan Ahmad, Kiflom Desta, Fabian Lang, and Dominik Schlechtweg. 2020. Shared task: Lexical semantic change detection in german. *CoRR*, arXiv:2001.07786.

N. Arefyev, M. Fedoseev, V. Protasov, D. Homskiy, A. Davletov, and A. Panchenko. 2021a. Deepmistake: Which senses are hard to distinguish for a word-in-context model. In *Computational linguistics and intellectual technologies*, 20, page 16 – 30, Russian Federation.

Nikolay Arefyev, Maksim Fedoseev, Vitaly Protasov, Daniil Homskiy, Adis Davletov, and Alexander Panchenko. 2021b. DeepMistake: Which senses are hard to distinguish for a word-in-context model. volume 2021-June, pages 16–30.

Carlos Santos Armendariz, Matthew Purver, Matej Ulčar, Senja Pollak, Nikola Ljubešić, and Mark Granroth-Wilding. 2020. CoSimLex: A resource for evaluating graded word similarity in context. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5878–5886, Marseille, France. European Language Resources Association.

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(1-3):89–113.

Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020. Overview of the EVALITA 2020 Diachronic Lexical Semantics (DIACR-Ita) Task. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Christin Beck. 2020. DiaSense at SemEval-2020 Task 1: Modeling sense change via pre-trained BERT embeddings. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

Felix Blessing, Johannes S. Sax, Julian Kaufmann, Wei Zhao, Nikolay Arefyev, and Dominik Schlechtweg. 2026. APODICTUS: Automatic Processing Of DICTionary Update candidateS. In *LREC*.

Pierluigi Cassotti, Pierpaolo Basile, and Nina Tahmasebi. 2024a. DWUGs-IT: Extending and standardizing lexical semantic change detection for Italian. In *Proceedings of the 10th Italian Conference on Computational Linguistics, Pisa, Italy, December 4 - December 6, 2024*, CEUR Workshop Proceedings. CEUR-WS.org.

Pierluigi Cassotti, Stefano De Pascale, and Nina Tahmasebi. 2024b. Using synchronic definitions and semantic relations to classify semantic change types. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4539–4553, Bangkok, Thailand. Association for Computational Linguistics.

Pierluigi Cassotti, Lucia Siciliani, Marco de Gemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. Xl-lexeme: Wic pretrained model for cross-lingual lexical semantic change. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Jing Chen, Emmanuele Chersoni, Dominik Schlechtweg, Jelena Prokic, and Chu-Ren Huang. 2023. ChiWUG: A graph-based evaluation dataset for Chinese lexical semantic change detection. In *Proceedings of the 4th International Workshop on Computational Approaches to Historical Language Change*, Singapore. Association for Computational Linguistics.

Tejaswi Choppa, Michael Roth, and Dominik Schlechtweg. 2025. Predicting median, disagreement and noise label in ordinal Word-in-Context data. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 65–77, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Paul Cook, Jey Han Lau, Diana McCarthy, and Timothy Baldwin. 2014. Novel word-sense identification. In *COLING*, pages 1624–1635. ACL.

Jader Martins Camboim de Sá, Marcos Da Silveira, and Cédric Pruski. 2025. Survey in characterization of semantic change. *Preprint*, arXiv:2402.19088.

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

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

Mariia Fedorova, Timothee Mickus, Niko Partanen, Janine Siewert, Elena Spaziani, and Andrey Kutuzov. 2024. AXOLOTL'24 shared task on multilingual explainable semantic change modeling. In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, pages 72–91, Bangkok, Thailand. Association for Computational Linguistics.

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.

Roksana Goworek and Haim Dubossarsky. 2026. Rethinking metrics for lexical semantic change detection. In *The Proceedings for the 6th International Workshop on Computational Approaches to Language Change (LChange'26)*, pages 147–161, Rabat, Morocco. 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.

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.

Daniil Homskiy and Nikolay Arefyev. 2022. DeepMistake at LSCDiscovery: Can a multilingual word-in-context model replace human annotators? In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 173–179, Dublin, Ireland. Association for Computational Linguistics.

Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882, Jeju Island, Korea. Association for Computational Linguistics.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*, 2(1):193–218.

Andres Karjus. 2025. Machine-assisted quantitizing designs: augmenting humanities and social sciences with artificial intelligence. *Humanities and Social Sciences Communications*, 12(1):1–18.

Adam Kilgarriff. 2007. *Word Senses*, chapter 2. Springer.

Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications.

Sinan Kurtyigit, Maike Park, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Lexical Semantic Change Discovery. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online. Association for Computational Linguistics.

Andrey Kutuzov and Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

Andrey Kutuzov and Lidia Pivovarova. 2021. Rushifteval: a shared task on semantic shift detection for russian. *Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog conference*.

Andrey Kutuzov, Samia Touileb, Petter Mæhlum, Tita Enstad, and Alexandra Wittemann. 2022. NorDiaChange: Diachronic semantic change dataset for Norwegian. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2563–2572, Marseille, France. European Language Resources Association.

Severin Laicher. 2020. Historical word sense clustering with deep contextualized word embeddings.

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.

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

Zhidong Ling, Taichi Aida, Teruaki Oka, and Mamoru Komachi. 2023. Construction of evaluation dataset for Japanese lexical semantic change detection. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 125–136, Hong Kong, China. Association for Computational Linguistics.

Daniel Loureiro, Aminette D'Souza, Areej Nasser Muhajab, Isabella A. White, Gabriel Wong, Luis Espinosa-Anke, Leonardo Neves, Francesco Barbieri, and Jose Camacho-Collados. 2022. TempoWiC: An evaluation benchmark for detecting meaning shift in social media. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3353–3359, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Suresh Manandhar and Ioannis Klapaftis. 2009. SemEval-2010 task 14: Evaluation setting for word sense induction & disambiguation systems. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 117–122, Boulder, Colorado. Association for Computational Linguistics.

Federico Martelli, Najla Kalach, Gabriele Tola, and Roberto Navigli. 2021. SemEval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (MCL-WiC). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 24–36, Online. Association for Computational Linguistics.

Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2020. Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France. European Language Resources Association.

Syrielle Montariol, Matej Martinc, and Lidia Pivovarova. 2021. Scalable and interpretable semantic change detection. In *2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Alp Mujko. 2026. Bridging the thresholding gap: Ordinal classification for word-in-context using cumulative link models. Master thesis, University of Stuttgart.

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

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. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Marko Pranjić, Kaja Dobrovoljc, Senja Pollak, and Matej Martinc. 2025. Tracking semantic change in slovene: A novel dataset and optimal transport-based distance. *Preprint*, arXiv:2402.16596.

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*, pages 1037–1047, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Dominik Schlechtweg, Pierluigi Cassotti, Bill Noble, David Alfter, Sabine Schulte im Walde, and Nina Tahmasebi. 2024a. More DWUGs: Extending and evaluating Word Usage Graph datasets in multiple languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14379–14393, Miami, Florida, USA. Association for Computational Linguistics.

Dominik Schlechtweg, Tejaswi Choppa, Wei Zhao, and Michael Roth. 2025. CoMeDi shared task: Median judgment classification & mean disagreement ranking with ordinal Word-in-Context judgments. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 33–47, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Dominik Schlechtweg, Anna Hätty, Marco del Tredici, and Sabine Schulte im Walde. 2019. A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy. 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 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.

Dominik Schlechtweg, Shafqat Mumtaz Virk, Pauline Sander, Emma Sköldberg, Lukas Theuer Linke, Tuo Zhang, Nina Tahmasebi, Jonas Kuhn, and Sabine Schulte im Walde. 2024b. 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.

Dominik Schlechtweg, Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Nikolay Arefyev. 2024c. Sense through time: Diachronic word sense annotations for word sense induction and lexical semantic change detection. *Language Resources and Evaluation*.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Emma Sköldberg, Shafqat Mumtaz Virk, Pauline Sander, Simon Hengchen, and Dominik Schlechtweg. 2024. Revealing semantic variation in Swedish using computational models of semantic proximity: Results from lexicographical experiments. In *Proceedings of the 21st EURALEX International Congress Lexicography and Semantics*.

Charles Spearman. 1904. The proof and measurement of association between two things. *American Journal of Psychology*, 15:88–103.

Sean Trott and Benjamin Bergen. 2021. RAW-C: Relatedness of ambiguous words in context (a new lexical resource for English). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7077–7087, Online. Association for Computational Linguistics.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Nash Whaley. 2025. Human and computational measurement of semantic relations. Master thesis, University of Stuttgart.

Sachin Yadav and Dominik Schlechtweg. 2025. XL-DURel: Finetuning sentence transformers for ordinal word-in-context classification. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 338–351, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. LSCDiscovery: A

shared task on semantic change discovery and detection in Spanish. In *Proceedings of the 3rd International Workshop on Computational Approaches to Historical Language Change*, Dublin, Ireland. Association for Computational Linguistics.

Frank D. Zamora-Reina, Felipe Bravo-Marquez, Dominik Schlechtweg, and Nikolay Arefyev. 2025. Can large language models compete with specialized models in lexical semantic change detection? In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*.

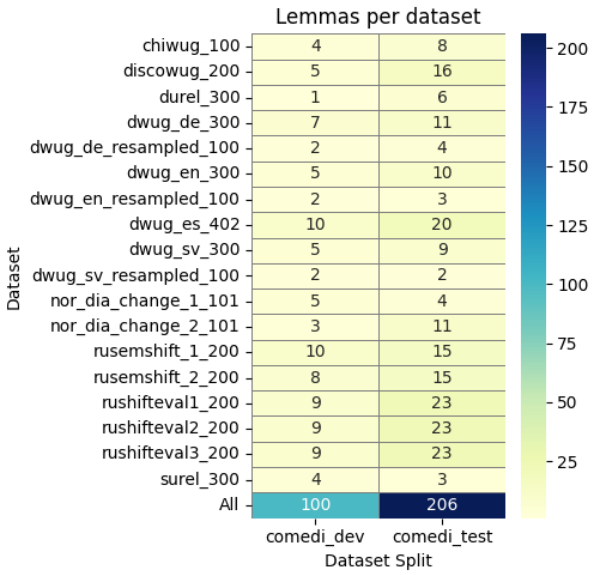

Figure 7: Number of target words in the CoMeDi split per dataset.

A CoMeDi Data Split

Find an overview of the number of target words in the dev and test portions of the CoMeDi data split in Figure 7.

B Specialized DeepMistake Models

For our main experiments we have selected the DeepMistake models that were trained on the general-purpose multilingual WiC dataset MCL-WiC (Martelli et al., 2021). However, the best models from (Arefyev et al., 2021a) and (Homskiy and Arefyev, 2022) were further fine-tuned on the training subsets of the corresponding LSCD shared tasks.

Table 2 compares the MCL model and the same model further fine-tuned on the RuSemShift dataset. We observe the largest improvements on the Russian datasets, and also surprisingly on the Spanish dataset in exchange to a big loss of performance on DURel. On other datasets the difference is 0.02 or smaller.

Table 3 further studies how the difference in these two WiC models affect their performance in LSCD. In these experiment we used the APD approach to LSCD, and averaged the WiC predictions for the COMPARE pairs left from the previous experiment, i.e. we used only the golden COMPARE pairs due to the limited computational budget. Similarly to the WiC task, we observe a big boost in LSCD performance on the Russian and Spanish datasets, but now the margins are larger. There are small or no changes on most other datasets. The exception this time is the resampled Swedish dataset where the ordering was mirrored, but this is likely due to chance because there are only 2 words in the CoMeDi test part of this dataset.

Dataset	MCL	MCL→RSS	Δ
chiwug_100	0.71	0.69	−0.02
discowug_200	0.62	0.61	−0.01
durel_300	0.67	0.61	−0.06
dwug_de_300	0.69	0.69	0.00
dwug_de_resampled_100	0.49	0.50	0.01
dwug_en_300	0.60	0.59	−0.02
dwug_en_resampled_100	0.52	0.49	−0.02
dwug_es_402	0.57	0.61	0.04
dwug_sv_300	0.59	0.59	0.00
dwug_sv_resampled_100	0.34	0.36	0.02
nor_dia_change_1_101	0.55	0.56	0.02
nor_dia_change_2_101	0.62	0.63	0.01
rushifteval1_200	0.48	0.54	0.06
rushifteval2_200	0.53	0.57	0.04
rushifteval3_200	0.56	0.60	0.04
surel_300	0.83	0.84	0.01

Table 2: Comparison of the general and specialized DeepMistake models on the WiC task. The results on RuSemShift are skipped because it was used as training and development data for the specialized model. Spearman's correlation with the human annotations when sorting pairs of usages is reported.

Dataset	MCL	MCL→RSS	Δ
chiwug_100	0.95	0.95	0.00
discowug_200	0.81	0.73	−0.08
durel_300	0.94	0.94	0.00
dwug_de_300	0.95	0.95	−0.01
dwug_de_resampled_100	0.80	0.80	0.00
dwug_en_300	0.95	0.90	−0.05
dwug_en_resampled_100	1.00	1.00	0.00
dwug_es_402	0.74	0.88	0.15
dwug_sv_300	0.90	0.88	−0.02
dwug_sv_resampled_100	−1.00	1.00	2.00
nor_dia_change_1_101	0.60	0.60	0.00
nor_dia_change_2_101	0.81	0.78	−0.03
rushifteval1_200	0.66	0.86	0.20
rushifteval2_200	0.53	0.74	0.21
rushifteval3_200	0.70	0.82	0.12
surel_300	1.00	1.00	0.00

Table 3: Comparison of the general and specialized DeepMistake models on the COMPARE task. Spearman's correlation with the ground truth ranking of words is reported.

To summarize, the results in this section once again confirm that fine-tuning on the target LSCD datasets can hugely boost the performance of WiC models for the LSCD task. Though the XL-DURel which was trained on all LSCD datasets generally outperforms other models across a wide range of LSCD datasets and languages, when focusing on a particular language training a specialized DeepMistake models for this language may give competitive results.

C Annotated Data Example

Find an example of an annotated Word Usage Graph from the DWUG EN dataset in Figure 8.

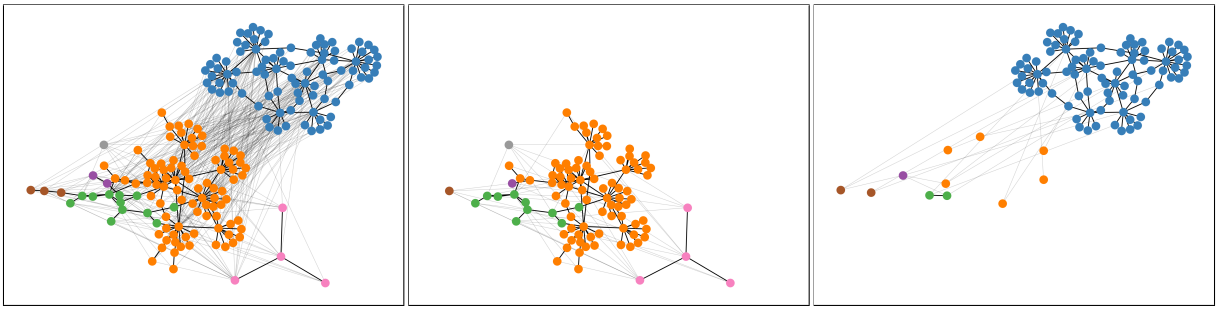

Figure 8: Word Usage Graph of English *plane* (left), subgraphs for 1st (middle) and 2nd time period (right).