University of Stuttgart
Germany

The LSCD Benchmark:
a Testbed for Diachronic Word Meaning Tasks

June 4, 2026

Dominik Schlechtweg[1], Sachin Yadav[1], Jonas Kuhn[1], Nikolay Arefyev[2]

University of Stuttgart[1], University of Oslo[2]

Introduction

- **Lexical Semantic Change Detection** (LSCD) (Schlechtweg, 2023)
  - goal: automate the analysis of changes in word meanings over time
    (1) *Der zweyte Theil vom Bauernrechte ist schon lange aus der **Presse**;*
    'The second part of Farmers' Rights already left the **press**;'
    (2) *Alle Freiheiten suspendirt! die persönliche Freiheit wie die der **Presse**!*
    'All freedoms suspended! the personal freedom as well as the one of the **press**!'
- complex, lemma-level task
- operationalized by **Word-in-Context** (WiC) annotations and subsequent **Word Sense Induction** (WSI) on weighted graph
- **modularity** and **heterogeneity** in models, datasets and tasks

→ create one repository[1] standardizing model component combinations, dataset preprocessing and evaluation

[1] `https://github.com/Garrafao/LSCDBenchmark`

Related Work

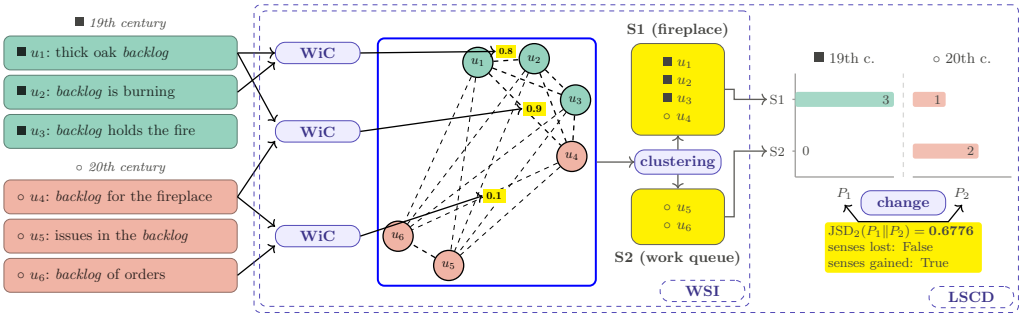
- shared tasks (Ahmad et al., 2020; Basile et al., 2020; Fedorova et al., 2024; Kutuzov & Pivovarova, 2021; Schlechtweg et al., 2020; Zamora-Reina et al., 2022)
- comprehensive studies with code repositories (Periti & Tahmasebi, 2024; Schlechtweg et al., 2019)

→ no comprehensive benchmark with flexible model & evaluation implementation

Tasks

1. Word-in-Context
2. Word Sense Induction
3. Lexical Semantic Change Detection

Tasks


19th century
u_1: thick oak backlog
u_2: backlog is burning
u_3: backlog holds the fire
20th century
u_4: backlog for the fireplace
u_5: issues in the backlog
u_6: backlog of orders
WiC
WiC
WiC
0.8
0.9
0.1
S1 (fireplace)
S2 (work queue)
clustering
WSI
19th c.
20th c.
S1
S2
3
1
0
2
P_1
change
P_2
$\mathrm{JSD}_2(P_1 \| P_2) = \mathbf{0.6776}$
senses lost: False
senses gained: True
LSCD

Datasets[2]

Dataset	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	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	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	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
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
...											
ChiWUG	ZH	40	10/22/8	40	4	61k	WiC, WSI, LSCD (B,G,C)	1954-1978	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. (2024)	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
...											

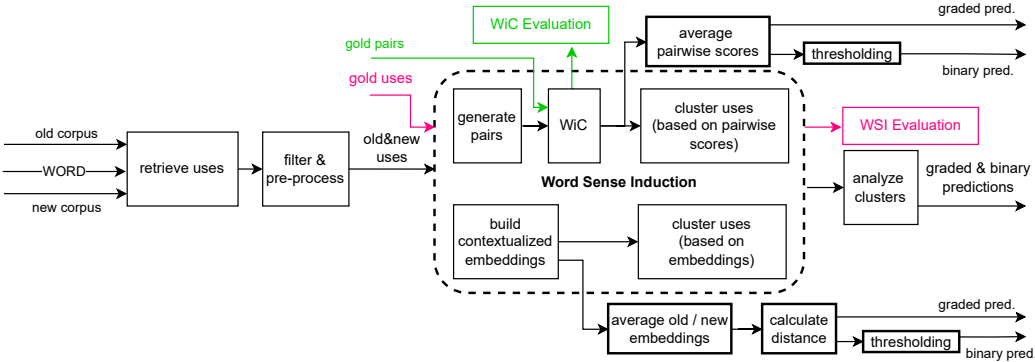
[2] https://www.ims.uni-stuttgart.de/data/wugs

Evaluation Procedures

- **WiC ranking** with Spearman's *rho*
- **WSI clustering** with Adjusted Rand Index
- **binary change classification** with F1
- **graded change ranking** with Spearman's *rho*
  - **COMPARE ranking** with Spearman's *rho*

- standard target word splits
  - CoMeDi split (Schlechtweg et al., 2025)

Models


old corpus
WORD
new corpus
retrieve uses
filter & pre-process
old&new uses
gold pairs
gold uses
WiC Evaluation
generate pairs
WiC
average pairwise scores
thresholding
graded pred.
binary pred.
cluster uses (based on pairwise scores)
WSI Evaluation
Word Sense Induction
analyze clusters
graded & binary predictions
build contextualized embeddings
cluster uses (based on embeddings)
average old / new embeddings
calculate distance
thresholding
graded pred.
binary pred.

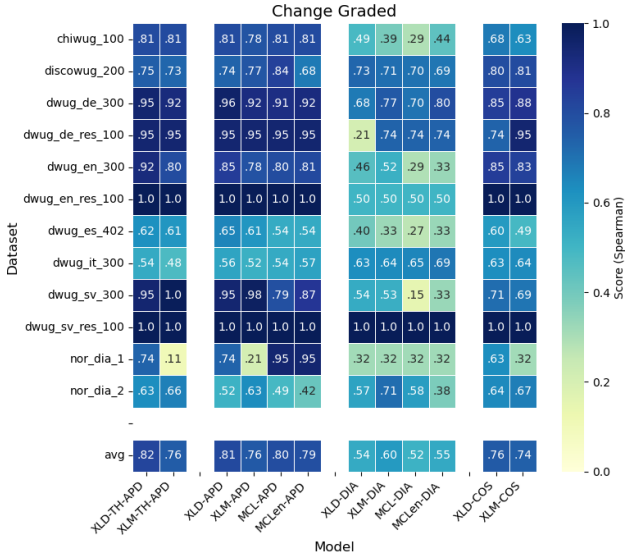
Usage Example

```
python main.py \
  dataset=dwug_de_210 \
  dataset/split=comedi_test \
  dataset/preprocessing=raw \
  task/wic@task.model.wic=contextual_embedder \
  task.model.wic.ckpt=bert-base-german-cased \
  task/wic/metric@task.model.wic.similarity_metric=cosine \
  task/lscd_graded@task.model=apd_compare_all \
  task=lscd_graded \
  evaluation=change_graded
```

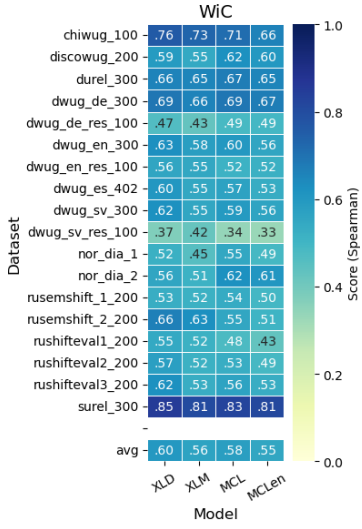
Experiments

- graded change ranking
- CoMeDi split
- WiC component
  - XL-LEXEME + cosine (Cassotti et al., 2023)
  - DeepMistake (MCL, MCLen) (Homskiy & Arefyev, 2022)
  - XL-DURel + cosine (Yadav & Schlechtweg, 2025)
- aggregation models
  - APD (Kutuzov & Giulianelli, 2020)
  - TH-APD (Schlechtweg et al., 2025)
  - DIA (Beck, 2020)
  - COS (Kutuzov & Giulianelli, 2020)

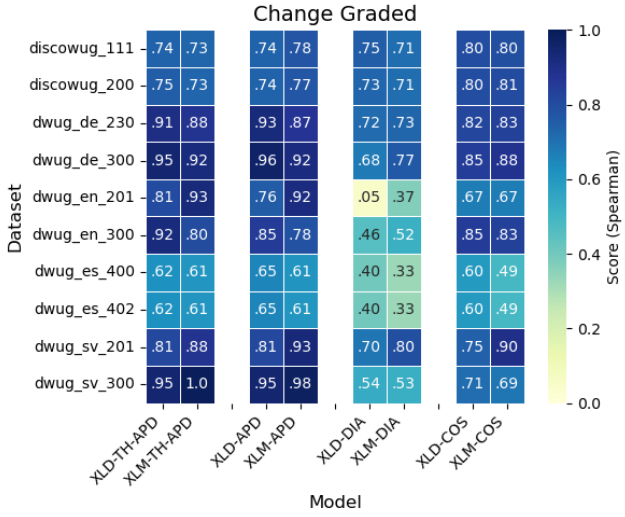
Which WiC model and which aggregate measure gives SOTA performance? Does WiC prediction discretization improve results?



Which model gives SOTA on diachronic WiC? Does WiC determine LSCD?



Are model performances reproducible with more reliable data? What is the performance development on incrementally annotated datasets?



Conclusion

- modular, flexible **benchmark repository**[3]
- integrates variety of models, datasets and evaluation procedures
- **findings**:
  - ordinal WiC models compete with their binary counterparts
  - discretization of graded WiC predictions helps
  - WiC performance roughly determines LSCD performance
  - performance on older dataset versions may be misleading

→ more closely **modelling the human annotation process** leads to better performance

[3] `https://github.com/Garrafao/LSCDBenchmark`

Limitations

- no clustering models
- non-optimal model training checkpoints
- data split size

Future Work

- clustering models (Schlechtweg et al., 2024)
- other training checkpoints (Homskiy & Arefyev, 2022)
- full or concatenated datasets
- recently developed aggregation measures (Goworek & Dubossarsky, 2026; Pranjić et al., 2025)

References I

Ahmad, A., Desta, K., Lang, F., & Schlechtweg, D. (2020). *Shared task: Lexical semantic change detection in german* (Vol. abs/2001.07786). Retrieved from https://arxiv.org/abs/2001.07786

Basile, P., Caputo, A., Caselli, T., Cassotti, P., & Varvara, R. (2020). Overview of the EVALITA 2020 Diachronic Lexical Semantics (DIACR-Ita) Task. In V. Basile, D. Croce, M. Di Maro, & L. C. Passaro (Eds.), *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*. Online: CEUR.org.

Beck, C. (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.

Cassotti, P., Basile, P., & Tahmasebi, N. (2024). 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-WS.org. Retrieved from https://clic2024.ilc.cnr.it/wp-content/uploads/2024/12/22_main_long.pdf

Cassotti, P., Siciliani, L., de Gemmis, M., Semeraro, G., & Basile, P. (2023, July). 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.

Chen, J., Chersoni, E., Schlechtweg, D., Prokic, J., & Huang, C.-R. (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. Retrieved from https://aclanthology.org/2023.lchange-1.10/

Fedorova, M., Mickus, T., Partanen, N., Siewert, J., Spaziani, E., & Kutuzov, A. (2024, aug). AXOLOTL'24 shared task on multilingual explainable semantic change modeling. In N. Tahmasebi et al. (Eds.), *Proceedings of the 5th workshop on computational approaches to historical language change* (pp. 72–91). Bangkok, Thailand: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.lchange-1.8 doi: 10.18653/v1/2024.lchange-1.8

Goworek, R., & Dubossarsky, H. (2026, mar). Rethinking metrics for lexical semantic change detection. In N. Tahmasebi et al. (Eds.), *The proceedings for the 6th international workshop on computational approaches to language change (LChange'26)* (pp. 147–161). Rabat, Morocco: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2026.lchange-1.13/ doi: 10.18653/v1/2026.lchange-1.13

Hätty, A., Schlechtweg, D., & Schulte im Walde, S. (2019). SURel: A gold standard for incorporating meaning shifts into term extraction. In *Proceedings of the 8th Joint Conference on Lexical and Computational Semantics* (pp. 1–8). Minneapolis, MN, USA. Retrieved from https://aclanthology.org/S19-1001/

References II

Homskiy, D., & Arefyev, N. (2022, May). DeepMistake at LSCDiscovery: Can a multilingual word-in-context model replace human annotators? In N. Tahmasebi, S. Montariol, A. Kutuzov, S. Hengchen, H. Dubossarsky, & L. Borin (Eds.), *Proceedings of the 3rd workshop on computational approaches to historical language change* (pp. 173–179). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.lchange-1.18 doi: 10.18653/v1/2022.lchange-1.18

Kutuzov, A., & Giulianelli, M. (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.

Kutuzov, A., & Pivovarova, L. (2021). Rushifteval: a shared task on semantic shift detection for russian. *Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog conference*.

Kutuzov, A., Touileb, S., Mæhlum, P., Enstad, T., & Wittemann, A. (2022, June). NorDiaChange: Diachronic semantic change dataset for Norwegian. In *Proceedings of the thirteenth language resources and evaluation conference* (pp. 2563–2572). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2022.lrec-1.274

Periti, F., & Tahmasebi, N. (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)* (pp. 4262–4282). Association for Computational Linguistics. Retrieved from http://dx.doi.org/10.18653/v1/2024.naacl-long.240 doi: 10.18653/v1/2024.naacl-long.240

Pranjić, M., Dobrovoljc, K., Pollak, S., & Martinc, M. (2025). *Tracking semantic change in slovene: A novel dataset and optimal transport-based distance*. Retrieved from https://arxiv.org/abs/2402.16596

Rodina, J., & Kutuzov, A. (2020, dec). RuSemShift: a dataset of historical lexical semantic change in Russian. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th international conference on computational linguistics* (pp. 1037–1047). Barcelona, Spain (Online): International Committee on Computational Linguistics. Retrieved from https://aclanthology.org/2020.coling-main.90 doi: 10.18653/v1/2020.coling-main.90

Schlechtweg, D. (2023). *Human and computational measurement of lexical semantic change*. Stuttgart, Germany. Retrieved from http://dx.doi.org/10.18419/opus-12833

Schlechtweg, D., Choppa, T., Zhao, W., & Roth, M. (2025, jan). CoMeDi shared task: Median judgment classification & mean disagreement ranking with ordinal Word-in-Context judgments. In M. Roth & D. Schlechtweg (Eds.), *Proceedings of context and meaning: Navigating disagreements in nlp annotation* (pp. 33–47). Abu Dhabi, UAE: International Committee on Computational Linguistics. Retrieved from https://aclanthology.org/2025.comedi-1.4/

References III

Schlechtweg, D., Hätty, A., del Tredici, M., & Schulte im Walde, S. (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* (pp. 732–746). Florence, Italy: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P19-1072/

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. (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. Retrieved from https://aclanthology.org/2020.semeval-1.1/

Schlechtweg, D., Schulte im Walde, S., & Eckmann, S. (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* (pp. 169–174). New Orleans, Louisiana. Retrieved from https://aclanthology.org/N18-2027/

Schlechtweg, D., Tahmasebi, N., Hengchen, S., Dubossarsky, H., & McGillivray, B. (2021, nov). 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* (pp. 7079–7091). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.567

Schlechtweg, D., Zamora-Reina, F. D., Bravo-Marquez, F., & Arefyev, N. (2024). Sense through time: Diachronic word sense annotations for word sense induction and lexical semantic change detection. *Language Resources and Evaluation*. Retrieved from https://doi.org/10.1007/s10579-024-09771-7

Yadav, S., & Schlechtweg, D. (2025, dec). XL-DURel: Finetuning sentence transformers for ordinal word-in-context classification. In K. Inui et al. (Eds.), *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* (pp. 338–351). Mumbai, India: The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. Retrieved from https://aclanthology.org/2025.findings-ijcnlp.19/

Zamora-Reina, F. D., Bravo-Marquez, F., & Schlechtweg, D. (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. Retrieved from https://aclanthology.org/2022.lchange-1.16/