University of Stuttgart
Germany

Models of **Word Usage Graphs** for **Lexical Semantic Change Detection** and Cognitive Insights into **Word Meaning**

March 20, 2025

Dominik Schlechtweg

Institute for Natural Language Processing, University of Stuttgart

Outline

Introduction

- Lexical Semantic Change Detection
  - **goal**: automate the analysis of changes in word meanings in text over time

(1) ***Mäuse*** *und Ratten sind selbstverständlich mit den europäischen Schiffen auch hierher gekommen.*
‘Of course, **mice** and rats also came here with the European ships.’

(2) *Deshalb eignet sich die* ***Maus*** *auch für verschiedene Betriebssysteme neben Windows und macOS.*
‘The **mouse** is therefore also suitable for various operating systems in addition to Windows and macOS.’

Two measurement paradigms

1. Word Usage Graphs (Schlechtweg, 2023)
    - compares **corpus** to **corpus**
    - builds on **Word Sense Induction** (Schütze, 1998)
2. Unrecorded Sense Detection (Erk, 2006; Fedorova et al., 2024)
    - compares **corpus** to **dictionary**
    - builds on **Word Sense Disambiguation** (Weaver, 1949/1955)

Word Usage Graphs

Human Measurement of Lexical Semantic Change

A	1824	and taking a knife from her pocket, she opened a vein in her little **arm**,	☺
B	1842	And those who remained at home had been heavily taxed to pay for the **arms**, ammunition;	⚔
C	1860	and though he saw her within reach of his **arm**, yet the light of her eyes seemed as far off	☺
		...	
D	1953	overlooking an **arm** of the sea which, at low tide, was a black and stinking mud-flat	🍶
E	1975	twelve miles of coastline lies in the southwest on the Gulf of Aqaba, an **arm** of the Red Sea.	🍶
F	1985	when the disembodied **arm** of the Statue of Liberty jets spectacularly out of the	☺

Table 1: Sample of diachronic corpus.

Word Use Pairs

(A) [...] and taking a knife from her pocket, she opened a vein in her little **arm**, and dipping a feather in the blood, wrote something on a piece of white cloth, which was spread before her. ☺

(D) It stood behind a high brick wall, its back windows overlooking an **arm** of the sea which, at low tide, was a black and stinking mud-flat [...]

Semantic Proximity Scale

4:	Identical
3:	Closely Related
2:	Distantly Related
1:	Unrelated

Table 2: DURel relatedness scale.

Graph representation

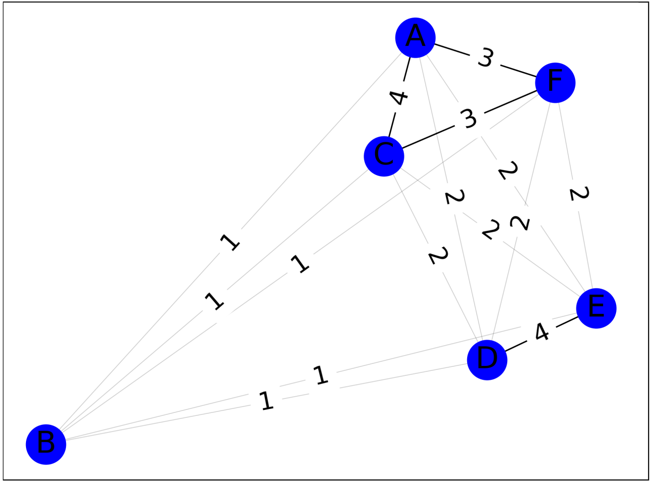


Figure 1: Word Usage Graph of English *arm*.

Clustering

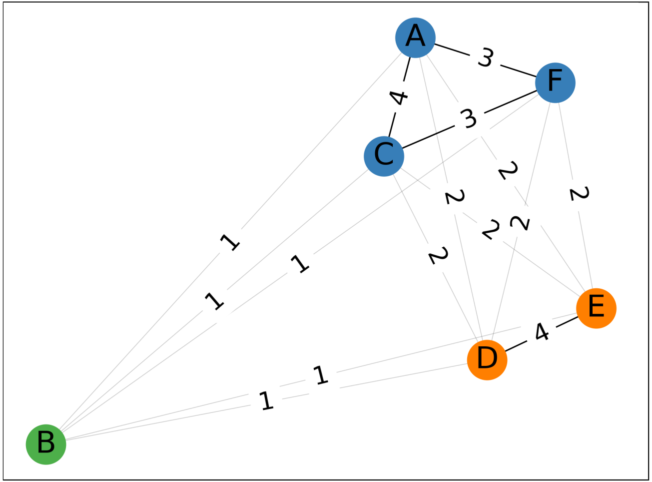
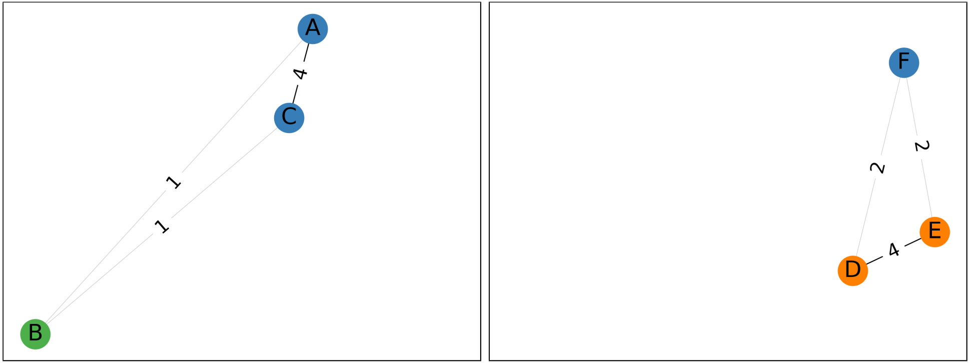


Figure 2: Word Usage Graph of English *arm*. $D = (3, 2, 1)$.

Lexical Semantic Change



$t_1,\ D_1 = (2, 0, 1)$

$t_2,\ D_2 = (1, 2, 0)$

Change Scores

- **binary change** (loss and gain of senses)
- **graded change** (changes in sense probabilities)

Evaluation Tasks

Task 1 **Binary classification**: for a set of target words, predict the binary change score

Task 2 **Ranking**: rank a set of target words according to their graded change score

(Schlechtweg et al., 2020)

Example: Swedish *ledning*[1]

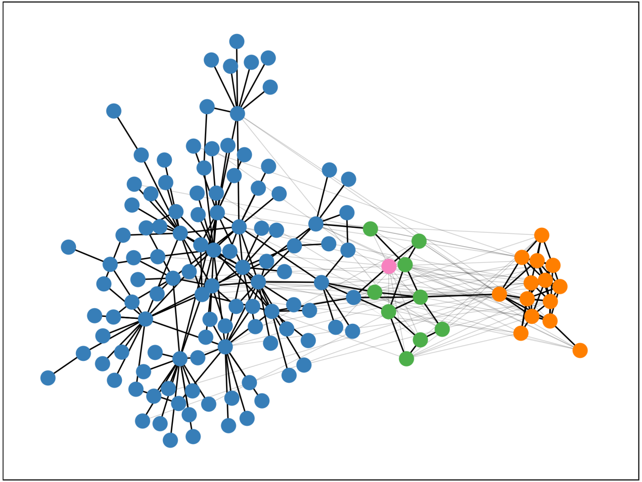


Figure 4: WUG of Swedish *ledning*.

[1] Datasets available at `https://www.ims.uni-stuttgart.de/data/wugs`

Example: Swedish *ledning*

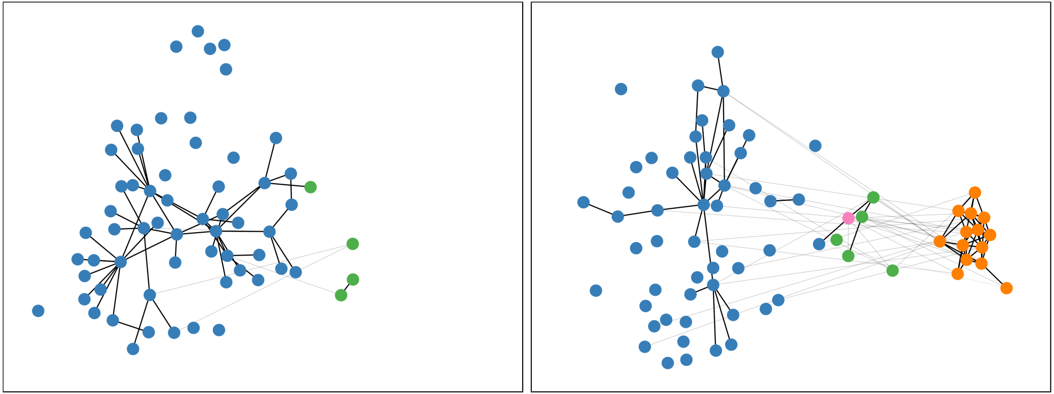


Figure 5: WUGs of Swedish *ledning*: subgraphs for 1st time period G_1 (left) and 2nd time period G_2 (right). $D_1 = (58, 0, 4, 0)$, $D_2 = (52, 14, 5, 1)$, $B(w) = 1$ and $G(w) = 0.34$.

Example: German *Eintagsfliege*

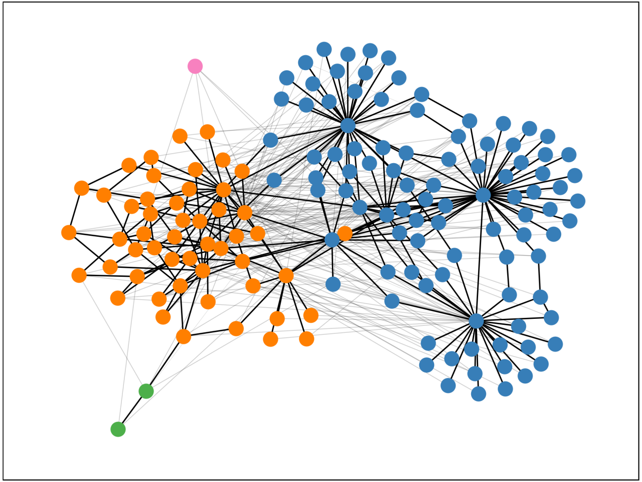


Figure 6: WUG of German *Eintagsfliege*.

Example: German *Eintagsfliege*

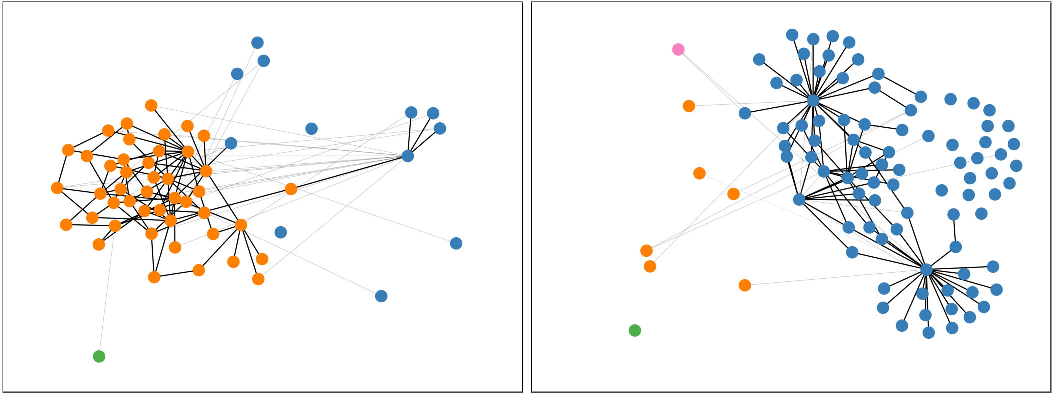


Figure 7: WUG of German *Eintagsfliege*: subgraphs for 1st time period G_1 (left) and 2nd time period G_2 (right). $D_1 = (12, 45, 0, 1)$, $D_2 = (85, 6, 1, 1)$, $B(w) = 0$ and $G(w) = 0.66$.

Summary of Annotation Steps

1. semantic proximity labeling
2. clustering
3. change measurement

Summary of Annotation Steps with Tasks

1. semantic proximity labeling ↔ **Word-in-Context Task**
2. clustering ↔ **Word Sense Induction**
3. change measurement ↔ **Lexical Semantic Change Detection** (including previous tasks)

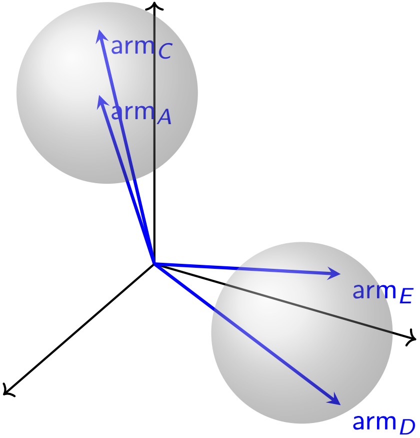
Computational Measurement of Lexical Semantic Change

- Typical **token**-based model is composed by
  1. semantic proximity model (e.g. similarity between contextualized embeddings)
  2. clustering method (optional)
  3. change measure

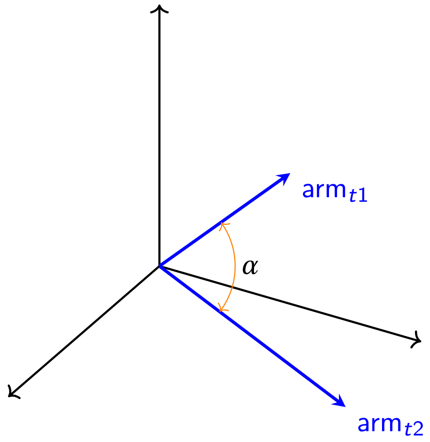
  - model the human measurement process
  - one vector per word use (BERT, ELMo)
- Typical **type**-based model is composed by
  1. semantic representation per word (type vector)
  2. alignment
  3. measure

  - do **not** model the human measurement process
  - one average vector per word (Word2Vec, GloVe)

Simple token-based Model



Simple type-based Model



SOTA Model Components

- SOTA Models used for the different levels:
  1. **semantic proximity**: DeepMistake, XL-Lexeme, GlossReader (Arefyev et al., 2021; Arefyev & Rachinskiy, 2021; Cassotti et al., 2023)
  2. **clustering**: Agglomerative, Spectral, Correlation, Stochastic Blockmodel (cf. Schlechtweg, Zamora-Reina, et al., 2024)
  3. **change measure**: Cluster gain/loss, Thresholding, Jensen Shannon Distance, Average Pairwise Distance (Kutuzov & Giulianelli, 2020; Lin, 1991)

Semantic Proximity Models

- aka **Word-in-Context** models (Pilehvar & Camacho-Collados, 2019)
- estimate degree of semantic proximity/same-sense probability for two input texts
- training data example:
  - (A) [. . .] and taking a knife from her pocket, she opened a vein in her little **arm**, and dipping a feather in the blood, wrote something on a piece of white cloth, which was spread before her. ☺
  - (D) It stood behind a high brick wall, its back windows overlooking an **arm** of the sea which, at low tide, was a black and stinking mud-flat [. . .]

    label: 0 (binary), 2 (ordinal)
- SOTA: DeepMistake, XL-Lexeme, GlossReader

  (Arefyev et al., 2021; Arefyev & Rachinskiy, 2021; Cassotti et al., 2023)
- optimized on binary multilingual semantic proximity data

  (Martelli et al., 2021; Pilehvar & Camacho-Collados, 2019; Raganato et al., 2020)
- have brought **major performance improvements** (Kutuzov & Pivovarova, 2021)
- can be mapped to ordinal labels (Schlechtweg et al., 2025)

Semantic Proximity Models

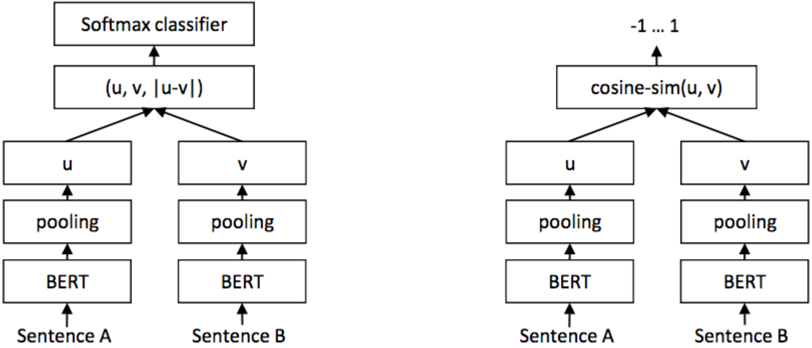


Figure 8: S-BERT (Reimers & Gurevych, 2019) training architecture used for XL-Lexeme.

Word Sense Induction Models

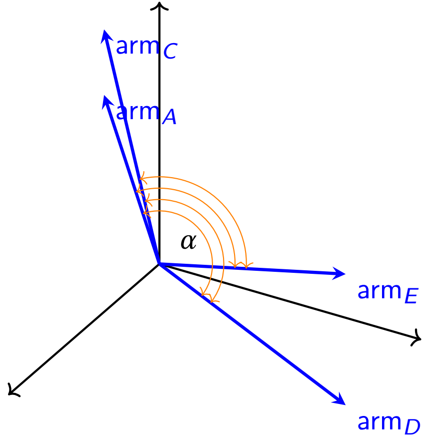
- aka **clustering**
- wealth of algorithms available
- Agglomerative, Spectral, Correlation, Stochastic Blockmodel

  (cf. Schlechtweg, Zamora-Reina, et al., 2024)
- **default**: Correlation Clustering (Bansal et al., 2004)
  - straightforward because used for ground truth clustering
  - **additional advantages**: finds number of clusters, intuitive

Change Measures

- **binary change**:
  - cluster gain/loss (cluster-based)
  - thresholding graded predictions (Schlechtweg et al., 2020)
- **graded change**:
  - Jensen Shannon Distance (cluster-based) (Lin, 1991)
  - Average Pairwise Distance (Kutuzov & Giulianelli, 2020)
- cluster-based vs. summary-based
- exact models vs. not
- summary-based dominates for graded change

SOTA Model for graded change: APD



Results

Lang.	Binary	Model	Graded	Model
Chinese			.73	XL-Lex.+APD
English	.70 (.67/.75)	BERT+HDBSCAN	.89	XL-Lex.+APD
German	.70 (.60/.82)	SGNS+thres.	.84	XL-Lex.+APD
Norwegian			.76	XL-Lex.+PRT
Russian			.86	XL-Lex.+APD
Swedish	.64 (.47/1.0)	XLM-R+K-means	.81	XL-Lex.+APD
Spanish	.72 (.62/.86)	GlossR.+APD+thres.	.74	GlossR.+APD

Table 3: SOTA performances on LSCD tasks (Cassotti et al., 2023; Periti & Tahmasebi, 2024a; Rachinskiy & Arefyev, 2022; Schlechtweg et al., 2020; Zamora-Reina et al., 2022). Values give F1 (P/R) for binary change and Spearman for graded change.

Summary

- **advantages**:
  - no need for sense definitions
  - rather explicit annotation criteria
- **disadvantages**:
  - questionable notion of semantic proximity
  - quadratically increased annotation load (Schlechtweg, Cassotti, et al., 2024)
  - need for clustering algorithm
- **open questions**:
  - clustering on gold data
  - cluster models & binary change (Graef, 2025)
  - application (Sköldberg et al., 2024)
  - multiple time periods (Periti & Tahmasebi, 2024b)
  - types of change (Whaley, 2024)
  - detect noisy usages (Choppa et al., 2025)
  - error analysis

Applications

- **DURel tool**: annotate, cluster and visualize WUGs with humans and computers[2] (Schlechtweg, Virk, et al., 2024)
- detect strongly changing words in German historical corpora with efficient **type-based** approaches (Kurtyigit et al., 2021)
- detect **unrecorded senses** in Swedish dictionary by comparing induced corpus sense number to dictionary sense number (Sander et al., 2024; Sköldberg et al., 2024)

[2] `https://durel.ims.uni-stuttgart.de/`

Clustering Word Usage Graphs

- WUGs need a clustering algorithm to **infer** senses
- default choice is **Correlation Clustering** (Schlechtweg et al., 2020)
  - motivated by **theory-driven** interpretation of annotation scale (Blank, 1997)
- **question**: can we evaluate this choice and do better?
- builds on previous ideas (Erk et al., 2013; McCarthy et al., 2016)

Graph representation

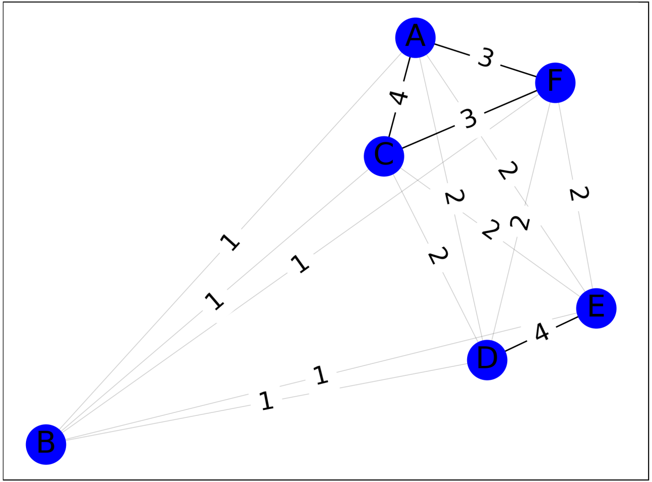


Figure 9: Word Usage Graph of English *arm*.

Clustering

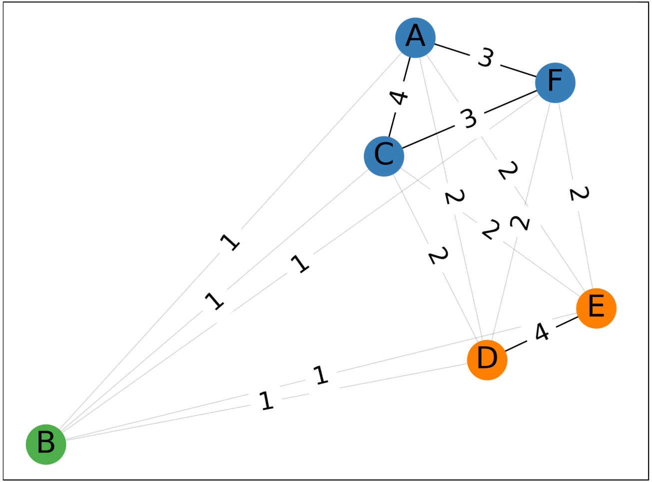


Figure 10: Word Usage Graph of English *arm*. $D = (3, 2, 1)$.

Problem

- **given**:
  - G = (U, E, W), weighted, undirected graph
  - nodes $u \in U$ represent word uses
  - weights $w \in W$ represent the human-annotated semantic proximity of a pair of uses (an edge) $(u_1, u_2) \in E$
- **task**:
  - cluster nodes $u \in U$ based on the edge weights such that uses with the same sense are in the same cluster

Data

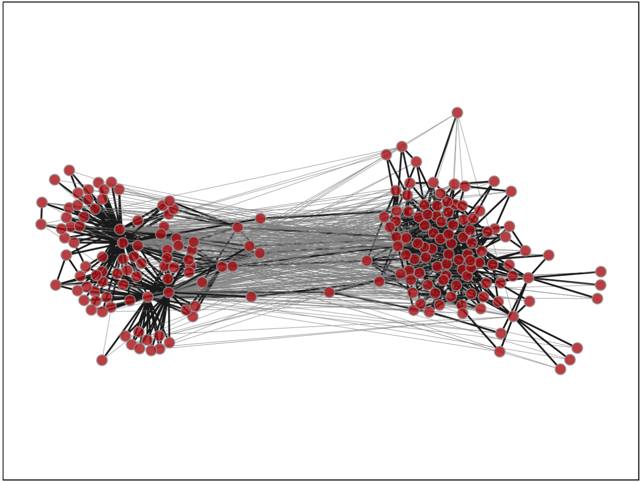


Figure 11: Word Usage Graph of German *zersetzen*.

Data

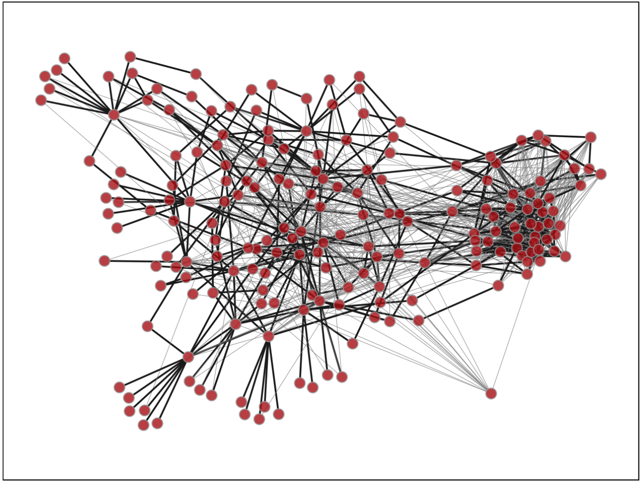


Figure 12: Word Usage Graph of German *Abgesang*.

Data[3]

	DWUG DE	DWUG DE Sense
n	50	24
N/V/A	34/14/2	16/7/1
|U|	≤100+≤100	25+25
AN	8	3
|J|	1.7	2.9
KRI	.67	.87
STYLE	use-use	use-sense

Table 4: Statistics for the latest version (V2.3.0) of DWUG DE and the DWUG DE Sense (V1.0.0) dataset. n = no. of target words, N/V/A = no. of nouns/verbs/adjectives, |U| = no. uses per word (t_1+t_2), AN = no. of annotators, |J| = avg. no. judgments per annotation instance, KRI = Krippendorff's α, STYLE = annotation style.

[3] Available at `https://www.ims.uni-stuttgart.de/data/wugs`

Models

Correlation Clustering (CC)

- $w \in W$ are shifted to obtain a set of **positive** and **negative** edges
- Let $C : U \mapsto L$ be some clustering on U
- $\phi_{E,C}$ is the set of positive (high) edges **across** any of the clusters in clustering C
- $\psi_{E,C}$ the set of negative (low) edges **within** any of the clusters
- correlation clustering searches for a clustering C that minimizes the sum of weighted cluster disagreements:

$$SWD(C) = \sum_{e \in \phi_{E,C}} W(e) + \sum_{e \in \psi_{E,C}} |W(e)| \, .$$

- **main assumption**:
  - weights above the threshold indicate same sense, below the threshold they indicate different sense

Weighted Stochastic Block Model (WSBM)

- a generative **probabilistic** model for random graphs (Aicher et al., 2014; Peixoto, 2019)
- popular in biology, physics and social sciences
- can be seen as an **explanation** of the data
- measures **uncertainty** over cluster assignments and allows for **model comparison**
- models nodes as part of blocks (clusters)
- **main assumption**:
  - nodes in the same block are **stochastically equivalent**, i.e., sampled from the same distribution

Inference of Block Structure

- we maximize the Bayesian posterior probability

$$P(b|A,x) = \frac{P(x|A,b)P(A|b)P(b)}{P(A,x)}$$

  where b is the inferred block structure, A is the (unweighted) observed graph, and x are the observed edge weights (Peixoto, 2017)
- approximation: multilevel agglomerative Markov chain Monte Carlo (Peixoto, 2014)

Evaluation

- leave-one-out cross-validation
- Adjusted Rand Index (ARI) (Hubert & Arabie, 1985)
  - accuracy on pairwise cluster agreements between nodes
  - controlled against agreement by chance

Results

method	ARI	t	dwn	t_{clps}	dist	mrg	dgr	weight	#folds
WSBM	.76	-	-	**False**	**binomial**	**True**	**True**	-	**20**
		-	-	2.4	binomial	True	False	-	2
		-	-	False	binomial	True	False	-	1
		-	-	2.3	binomial	True	True	-	1
CC	.72	**2.5**	**True**	**2.3**	-	-	-	-	**18**
		2.4	True	2.4	-	-	-	-	2
		2.5	True	2.4	-	-	-	-	1
		2.6	False	2.3	-	-	-	-	1
		2.9	True	2.3	-	-	-	-	1
		2.6	True	2.3	-	-	-	-	1

Table 5: The configurations of hyperparameters selected for each method in at least one CV fold. The configuration selected for the majority of folds is in **bold**. "-" marks non-applicable parameters for the respective method.

Conclusion

- we inferred sense structure in WUGs exploiting patterns of semantic proximity
- the probabilistic model outperformed heuristic model
- has additional **advantages**:
  - model selection allows **principled** inference of sense structure
  - **rigorous** comparison to other probabilistic models (Duda & Hart, 1973; Hoff et al., 2002)
  - inferred models can be used for **simulation** of realistic WUGs
- **future work**:
  - what do model assumptions imply?
  - reproducibility?
  - compare different types of probabilistic models?

Testing Cognitive Hypotheses in Word Usage Graphs

- Bayesian models of graphs allow to compare the **plausibility** of different models, given the observed data
- can be seen as different **explanations** of the data
- models make different assumptions on blocks, semantic proximity and their relations
- **assumption**: different cognitive organization of lexical information may imply different block structures
- **idea**: by selecting the most plausible block model we can learn about the cognitive organization of lexical information

Testing Cognitive Hypotheses in Word Usage Graphs

- **some questions**:
  - do senses overlap? (Airoldi et al., 2008; Peixoto, 2015)
  - is semantic proximity sampled from a latent (semantic) space? (Erk et al., 2013; Hoff et al., 2002)
  - does (one) semantic proximity exist?
  - should annotators be modeled individually? (Peixoto, 2017)
  - how to model ambiguity and disagreement? (Schlechtweg et al., 2025)

Data

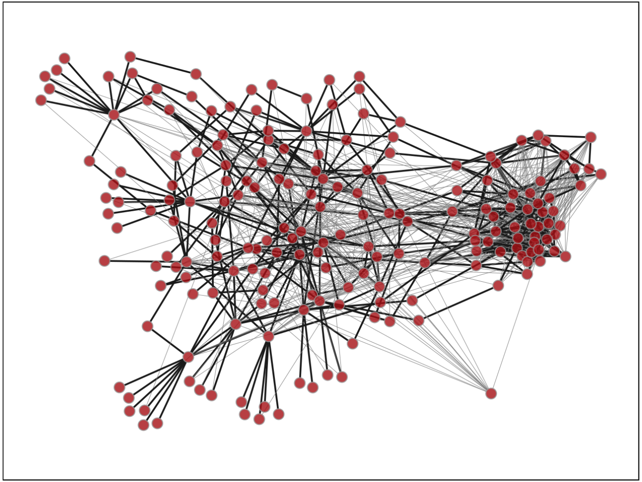


Figure 13: Word Usage Graph of German *Abgesang*.

References I

Aicher, C., Jacobs, A. Z., & Clauset, A. (2014, Jun). Learning latent block structure in weighted networks. *Journal of Complex Networks*, *3*(2), 221—248. Retrieved from http://dx.doi.org/10.1093/comnet/cnu026 doi: 10.1093/comnet/cnu026

Airoldi, E. M., Blei, D. M., Fienberg, S. E., & Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, *9*, 1981–2014.

Arefyev, N., Fedoseev, M., Protasov, V., Homskiy, D., Davletov, A., & Panchenko, A. (2021). DeepMistake: Which senses are hard to distinguish for a word-in-context model. In (Vol. 2021-June, pp. 16–30). Retrieved from https://www.dialog-21.ru/media/5491/arefyevnplusetal133.pdf

Arefyev, N., & Rachinskiy, M. (2021). Zero-shot cross-lingual transfer of a gloss language model for semantic change detection. In (Vol. 2021-June, pp. 578–586). doi: 10.28995/2075-7182-2021-20-578-586

Bansal, N., Blum, A., & Chawla, S. (2004). Correlation clustering. *Machine Learning*, *56*(1-3), 89–113. doi: 10.1023/B:MACH.0000033116.57574.95

Blank, A. (1997). *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen*. Tübingen: Niemeyer.

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.

Choppa, T., Roth, M., & Schlechtweg, D. (2025, jan). Predicting median, disagreement and noise label in ordinal word-in-context data. In M. Roth & D. Schlechtweg (Eds.), *Proceedings of context and meaning: Navigating disagreements in nlp annotation* (pp. 65–77). Abu Dhabi, UAE: International Committee on Computational Linguistics. Retrieved from https://aclanthology.org/2025.comedi-1.6/

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*.

Erk, K. (2006, June). Unknown word sense detection as outlier detection. In R. C. Moore, J. Bilmes, J. Chu-Carroll, & M. Sanderson (Eds.), *Proceedings of the human language technology conference of the NAACL, main conference* (pp. 128–135). New York City, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N06-1017

Erk, K., McCarthy, D., & Gaylord, N. (2013). Measuring word meaning in context. *Computational Linguistics*, *39*(3), 511–554.

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

Graef, L. (2025). *Erkennung binärer lexikalisch-semantischer veränderungen* (Bachelor thesis). University of Stuttgart.

References II

Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, *97*(460), 1090–1098. doi: 10.1198/016214502388618906

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218.

Kurtyigit, S., Park, M., Schlechtweg, D., Kuhn, J., & Schulte im Walde, S. (2021, aug). 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. Retrieved from `https://aclanthology.org/2021.acl-long.543/`

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*.

Lin, J. (1991, 01). Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, *37*, 145–151.

Martelli, F., Kalach, N., Tola, G., & Navigli, R. (2021, aug). SemEval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (MCL-WiC). In A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, & X. Zhu (Eds.), *Proceedings of the 15th international workshop on semantic evaluation (semeval-2021)* (pp. 24–36). Online: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2021.semeval-1.3` doi: 10.18653/v1/2021.semeval-1.3

McCarthy, D., Apidianaki, M., & Erk, K. (2016). Word sense clustering and clusterability. *Computational Linguistics*, *42*(2), 245–275.

Peixoto, T. P. (2014, Jan). Efficient monte carlo and greedy heuristic for the inference of stochastic block models. *Physical Review E*, *89*(1). Retrieved from `http://dx.doi.org/10.1103/PhysRevE.89.012804` doi: 10.1103/physreve.89.012804

Peixoto, T. P. (2015, Mar). Model selection and hypothesis testing for large-scale network models with overlapping groups. *Physical Review X*, *5*, 011033. doi: 10.1103/PhysRevX.5.011033

Peixoto, T. P. (2017, 08). Nonparametric weighted stochastic block models. *Physical Review E*, *97*. doi: 10.1103/PhysRevE.97.012306

Peixoto, T. P. (2019). Bayesian stochastic blockmodeling. In *Advances in network clustering and blockmodeling* (p. 289-332). John Wiley & Sons, Ltd. doi: 10.1002/9781119483298.ch11

References III

Periti, F., & Tahmasebi, N. (2024a). 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

Periti, F., & Tahmasebi, N. (2024b, August). Towards a complete solution to lexical semantic change: an extension to multiple time periods and diachronic word sense induction. In N. Tahmasebi et al. (Eds.), *Proceedings of the 5th workshop on computational approaches to historical language change* (pp. 108–119). Bangkok, Thailand: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.lchange-1.10/ doi: 10.18653/v1/2024.lchange-1.10

Pilehvar, M. T., & Camacho-Collados, J. (2019, June). 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)* (pp. 1267–1273). Minneapolis, Minnesota: Association for Computational Linguistics. doi: 10.18653/v1/N19-1128

Rachinskiy, M., & Arefyev, N. (2022). GlossReader at LSCDiscovery: Train to select a proper gloss in english – discover lexical semantic change in spanish. In *Proceedings of the 3rd international workshop on computational approaches to historical language change*. Dublin, Ireland: Association for Computational Linguistics.

Raganato, A., Pasini, T., Camacho-Collados, J., & Pilehvar, M. T. (2020, nov). XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 7193–7206). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-main.584 doi: 10.18653/v1/2020.emnlp-main.584

Reimers, N., & Gurevych, I. (2019, nov). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 3982–3992). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1410 doi: 10.18653/v1/D19-1410

Sander, P., Hengchen, S., Zhao, W., Ma, X., Sköldberg, E., Virk, S. M., & Schlechtweg, D. (2024). The DURel Annotation Tool: Using fine-tuned LLMs to discover non-recorded senses in multiple languages. In *Proceedings of the Workshop on Large Language Models and Lexicography at 21st EURALEX International Congress Lexicography and Semantics.* Retrieved from https://www.cjvt.si/wp-content/uploads/2024/10/LLM-Lex_2024_Book-of-Abstracts.pdf

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

References IV

Schlechtweg, D., Cassotti, P., Noble, B., Alfter, D., Schulte Im Walde, S., & Tahmasebi, N. (2024, nov). More DWUGs: Extending and evaluating word usage graph datasets in multiple languages. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 14379–14393). Miami, Florida, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.emnlp-main.796

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/

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://www.aclweb.org/anthology/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://www.aclweb.org/anthology/N18-2027/

Schlechtweg, D., Virk, S. M., Sander, P., Sköldberg, E., Theuer Linke, L., Zhang, T., . . . Schulte im Walde, S. (2024, mar). The DURel annotation tool: Human and computational measurement of semantic proximity, sense clusters and semantic change. In N. Aletras & O. De Clercq (Eds.), *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: System demonstrations* (pp. 137–149). St. Julians, Malta: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.eacl-demo.15

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

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

Sköldberg, E., Virk, S. M., Sander, P., Hengchen, S., & Schlechtweg, D. (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.* Retrieved from https://euralex.org/publications/revealing-semantic-variation-in-swedish-using-computational-models-of-semantic-proximity-results-from-lexicographical-experiments/

References V

Weaver, W. (1949/1955). Translation. In W. N. Locke & A. D. Boothe (Eds.), *Machine translation of languages* (pp. 15–23). Cambridge, MA: MIT Press. (Reprinted from a memorandum written by Weaver in 1949.)

Whaley, N. (2024). *Human and computational measurement of semantic relations* (Master thesis). University of Stuttgart.

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/

Appendix: Annotation Scale

4: Identical	Identity
3: Closely Related	Context Variance
2: Distantly Related	Polysemy
1: Unrelated	Homonymy

Table 6: The DURel relatedness scale (Schlechtweg et al., 2018) on the left and its interpretation from Schlechtweg (2023, p. 33) on the right.