University of Stuttgart
Germany

CHANGE
IS KEY

More DWUGs: Extending and Evaluating Word Usage Graph Datasets in Multiple Languages

October 29, 2024

Dominik Schlechtweg[1], Pierluigi Cassotti[2], Bill Noble[2], David Alfter[2], Sabine Schulte im Walde[1], Nina Tahmasebi[2]

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

UNIVERSITY OF GOTHENBURG

Introduction

- **Word Usage Graphs (WUGs)**: a new word sense annotation paradigm (Schlechtweg et al., 2020, 2021)
  - humans provide semantic proximity judgments of pairs of word uses
    - → represented in a weighted graph
    - → clustered with a graph clustering algorithm
  - avoids the need for a sense inventory
- **problems**:
  - annotation load
  - validity
  - robustness
  - replicability
- **aim**: quantify the problems and improve the data
- **approach**:
  - add additional rounds of annotation
  - compare against an external gold standard
  - resample and re-annotate previous data

A WUG example

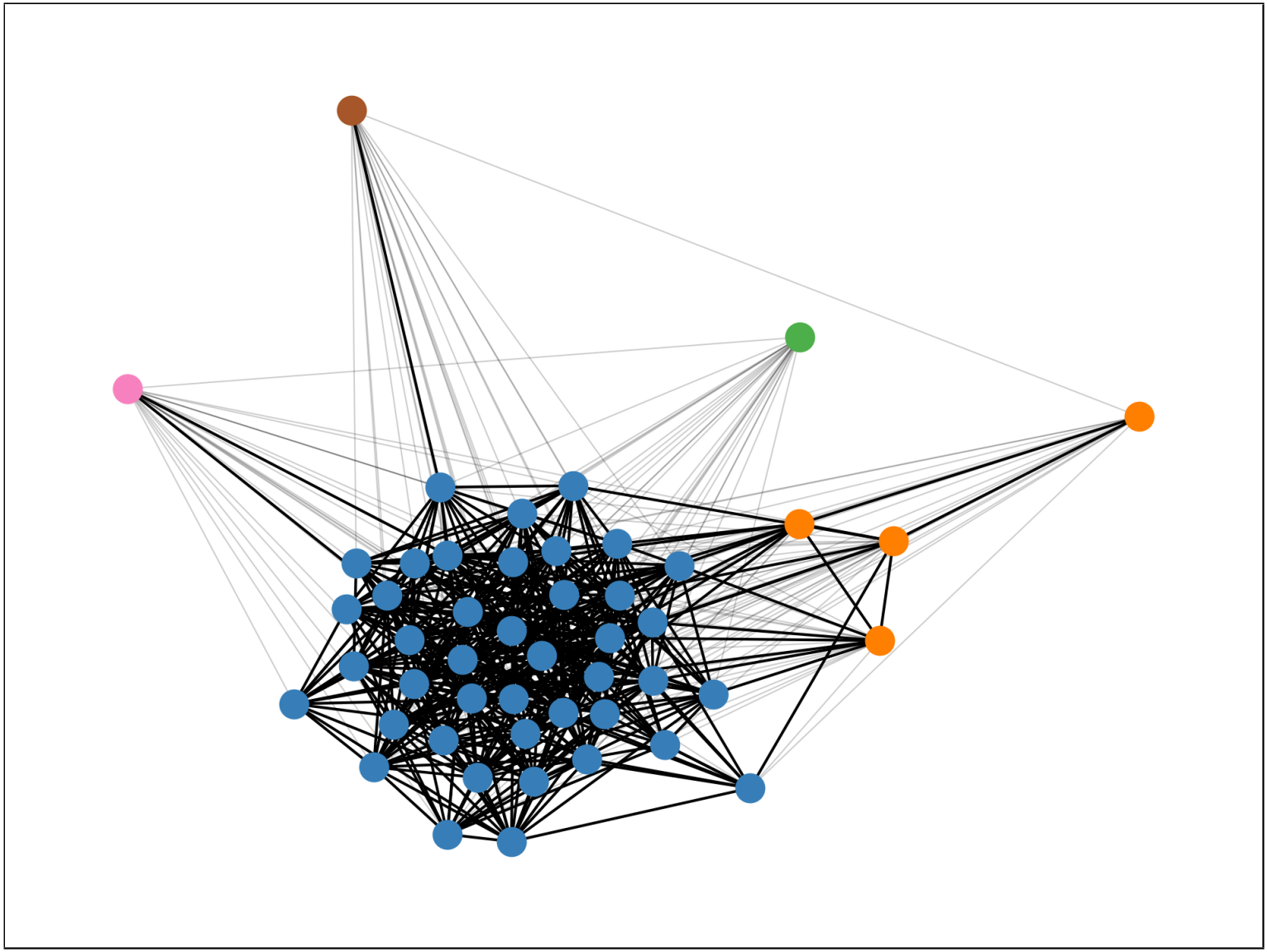


Figure 1: WUG of German *anpflanzen*.

Related work

- three established word sense annotation procedures

(Erk et al., 2013)

1. **use-sense**

   use: [...] taking a knife from her pocket, she opened a vein in her little **arm**.
   sense1: a human limb
   sense2: weapon system

2. **lexical substitution**

   use: And those who remained at home had been heavily taxed to pay for the **arms**, ammunition; fortifications, and all the other endless expenses of a war.

3. **use-use**

   use1: [...] taking a knife from her pocket, she opened a vein in her little **arm**.
   use2: It stood behind a high brick wall, its back windows overlooking an **arm** of the sea.

Corpus

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

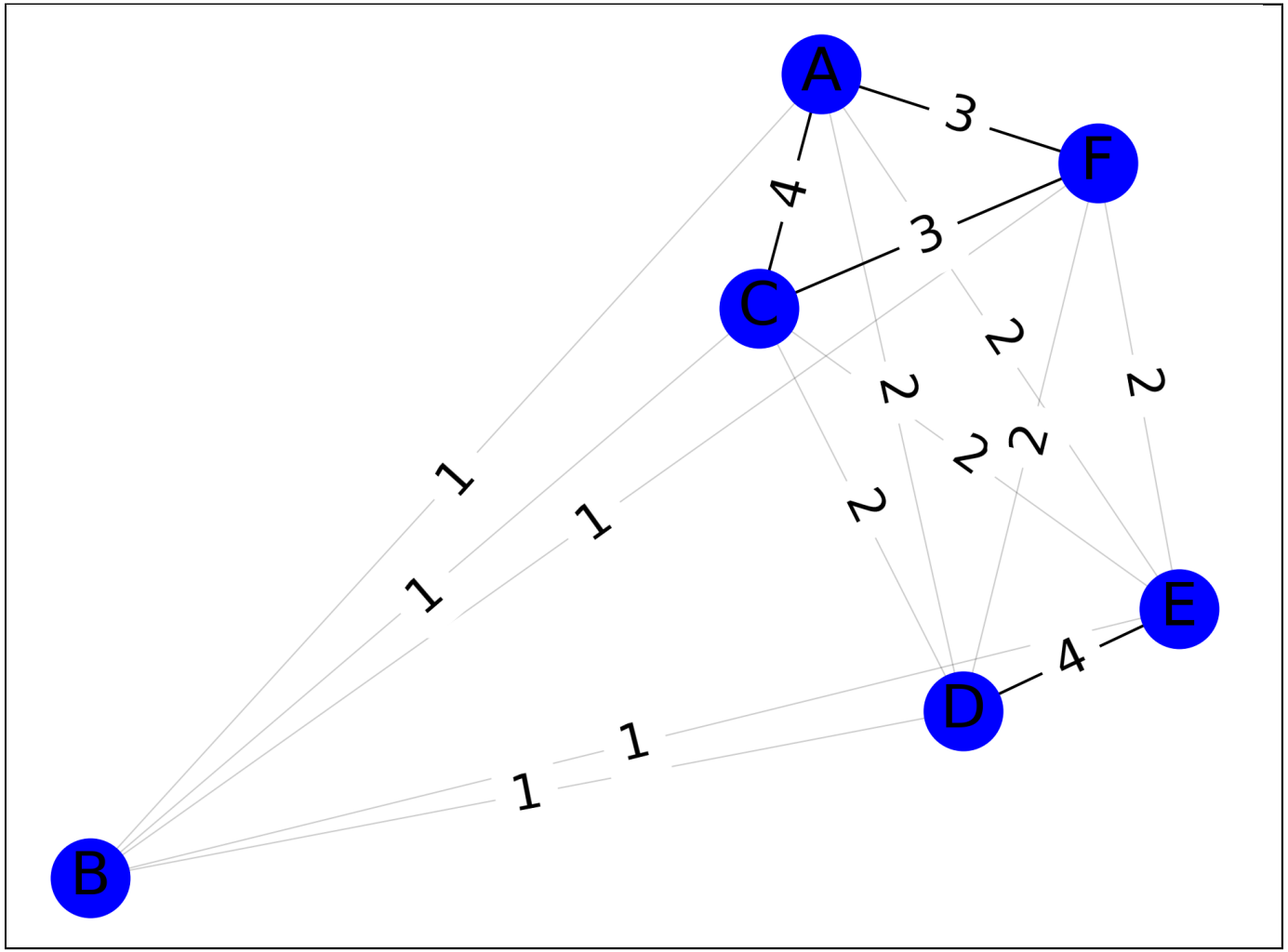


Figure 2: Word Usage Graph of English *arm*.

Clustering

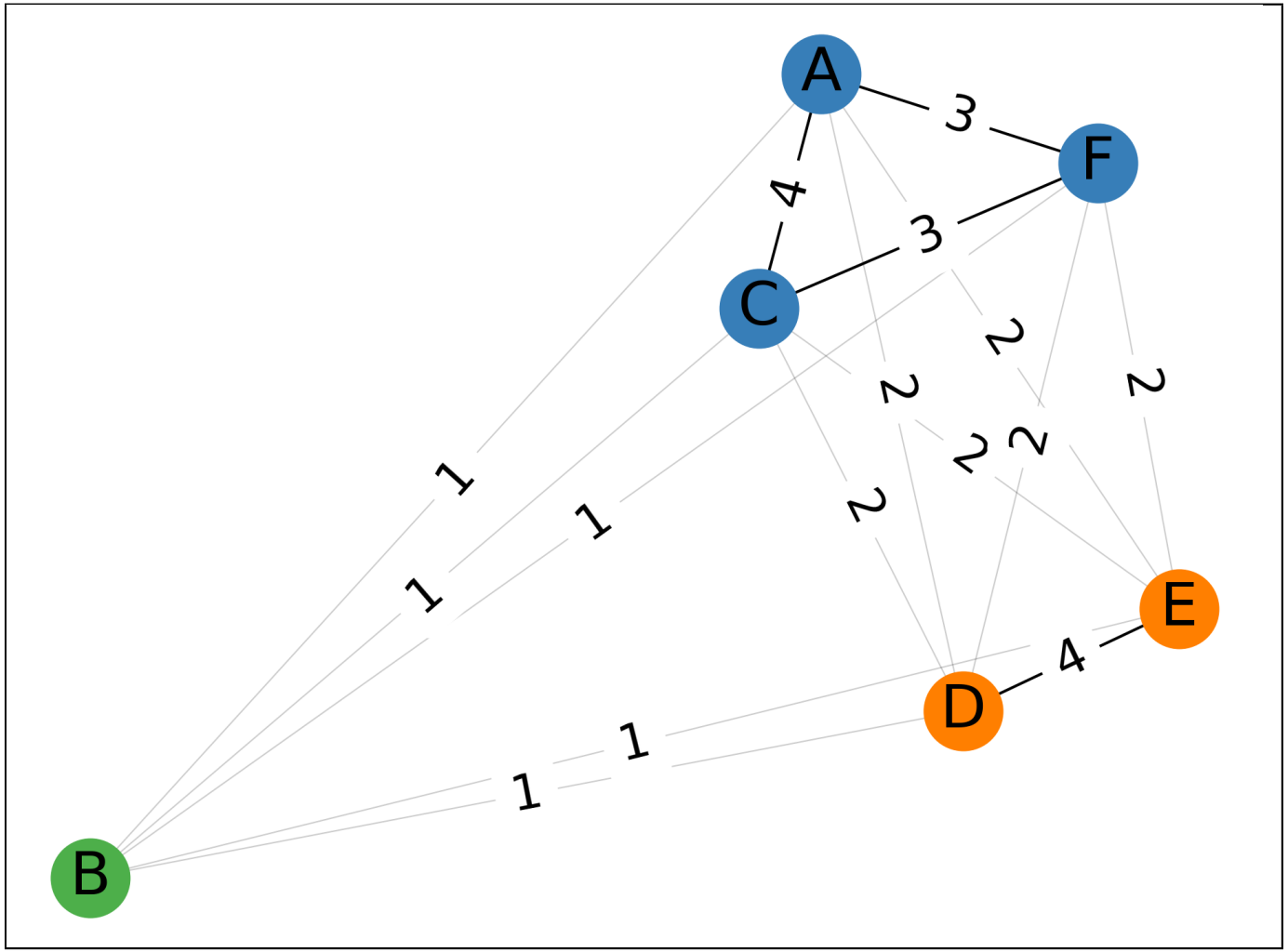
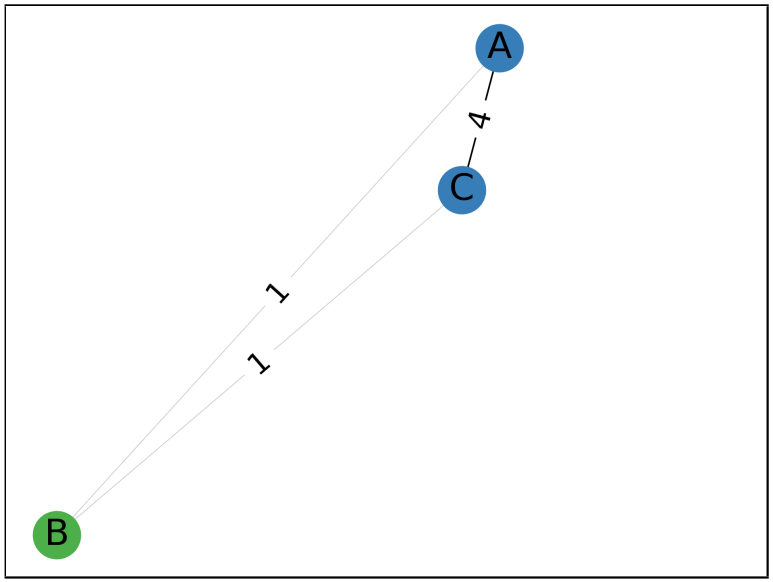
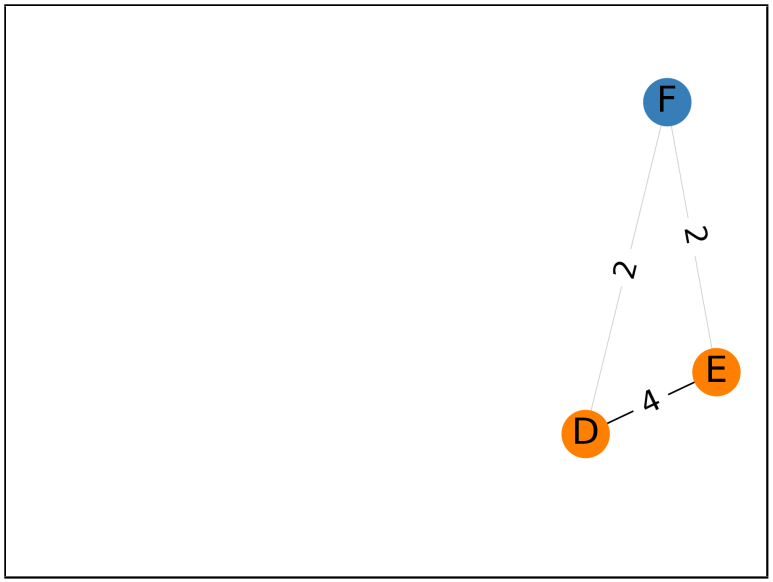


Figure 3: Word Usage Graph of English *arm*. $D = (3, 2, 1)$.

Lexical Semantic Change



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



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

Example: Swedish *ledning*[1]

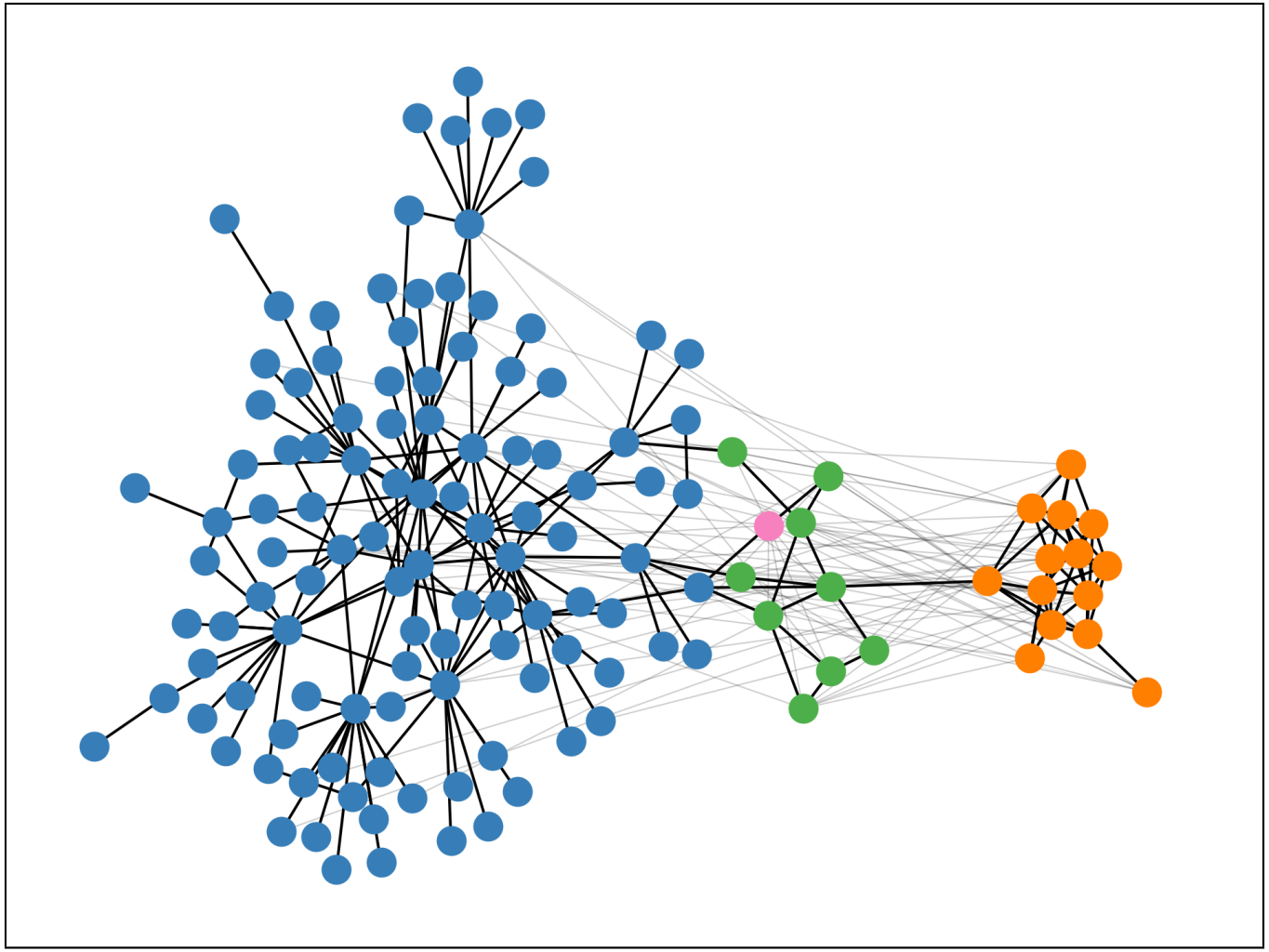


Figure 5: WUG of Swedish *ledning*.

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

Summary of Annotation Steps

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

Data

- **DWUG** (Schlechtweg et al., 2021)
  - English, German, Swedish
  - widely used
  - **many uses** per word (≤ 200)
  - sophisticated edge sampling algorithm
  - annotated in multiple rounds
  - very **sparsely annotated**
  - many small clusters are not connected
- **DiscoWUG** (Kurtyigit et al., 2021)
  - German
  - extends DWUG
  - **few uses** per word (50)
  - simple random edge sampling
  - annotated in one round
  - rather **densely annotated**
  - only few small clusters are not connected

Data

- **DWUG DE Sense** (Schlechtweg et al., 2024)
  - German
  - re-annotates subset of DWUG DE in classical use-sense style
  - few uses per word (50)
  - cleaned on the use level
  - serves as **gold standard** for comparison

Annotation

- **DWUG**
  - add **two more rounds**
  - random sampling + connecting clusters
- **DiscoWUG**
  - add **one more round**
  - connecting clusters
- **DWUG resampled**
  - **resample** uses for DWUG datasets
  - 15 words per language
  - 50 uses per word
  - random sampling

Result

	|E|			+|J|			|J|		
	DE	EN	SV	DE	EN	SV	DE	EN	SV
1–4	2.75	2.48	2.15	74	69	48	40K	36K	24K
1–5	3.39	3.18	2.73	142	193	191	48K	46K	37K
1–6	**4.90**	**5.09**	**4.61**	**297**	**487**	**394**	**63K**	**69K**	**55K**
resampled	44.75	35.29	59.85				10K	7K	16K

Table 3: Coverage by annotation round for DWUG datasets. |E|: average percentage of annotated edges, +|J|: average increase in number of judgments per word. |J|: absolute number of judgments.

Experiments

- we evaluate
  - the **validity** of the inferred clusters over rounds of annotation by comparing them to an external gold standard
  - the **robustness** of the final clusterings by perturbing the graphs with random annotations
  - their **replicability** through a complete resampling and re-annotation of data
- Adjusted Rand Index (ARI)
- Jensen Shannon distance (JSD)

Validity of clusters

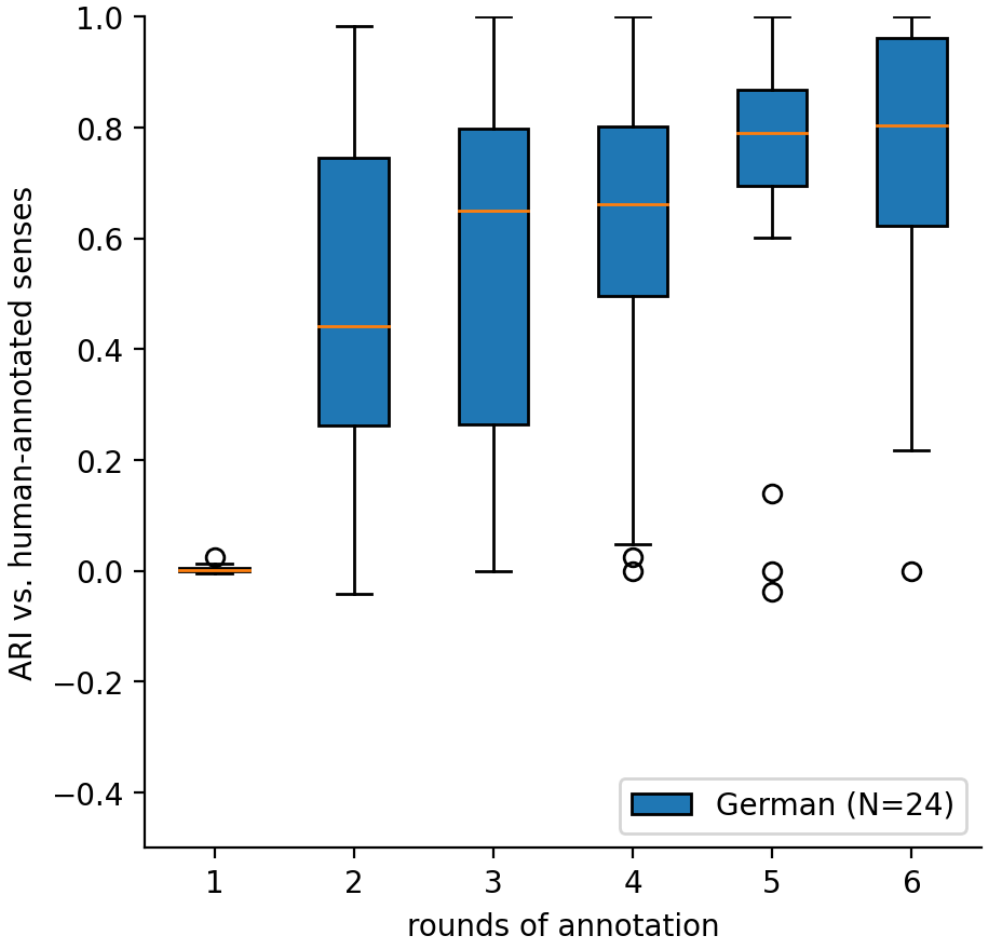


Figure 6: ARI of DWUG DE clusters over rounds vs. DWUG DE Sense annotation.

Validity of clusters

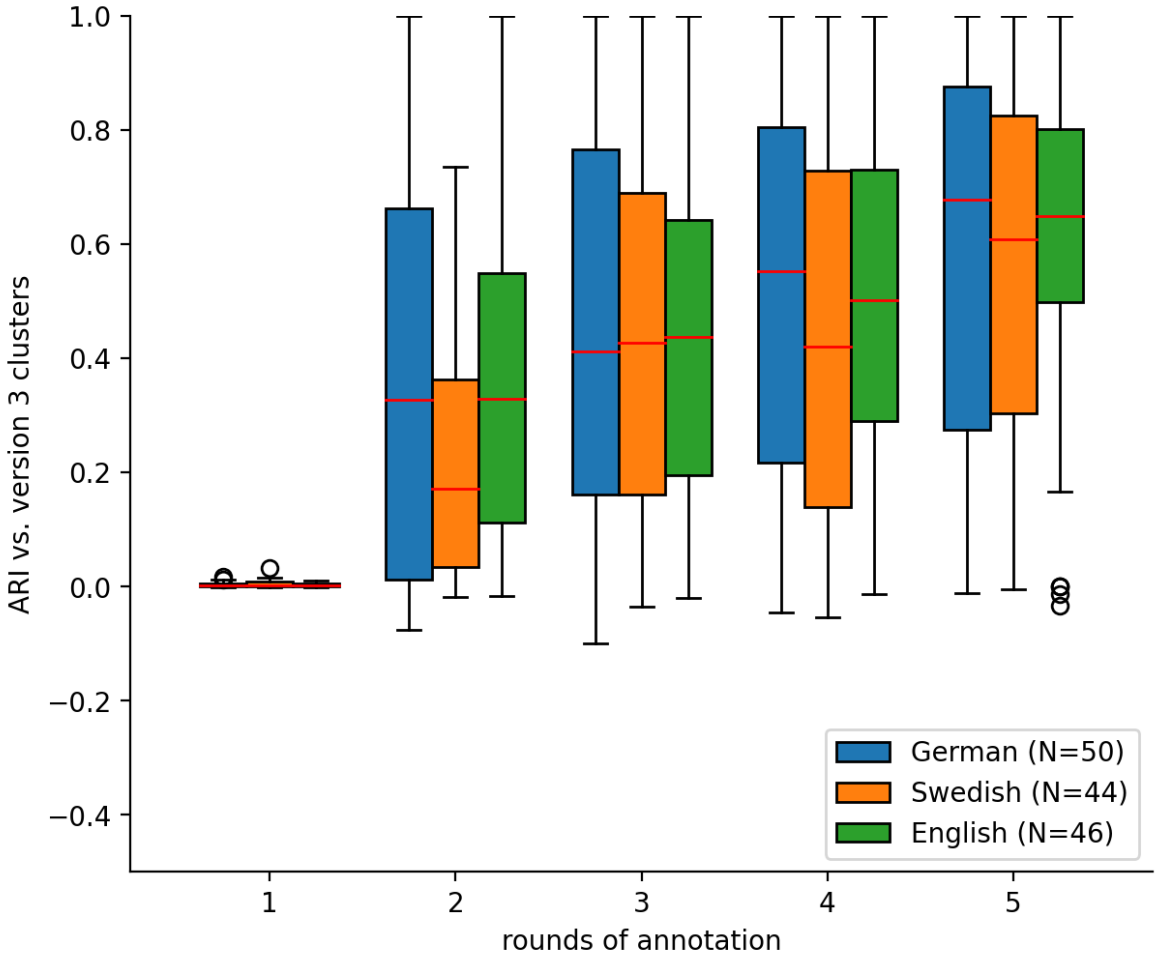


Figure 7: ARI of DWUG DE/EN/SV clusters over rounds vs. full data set (last round).

Robustness of clusters

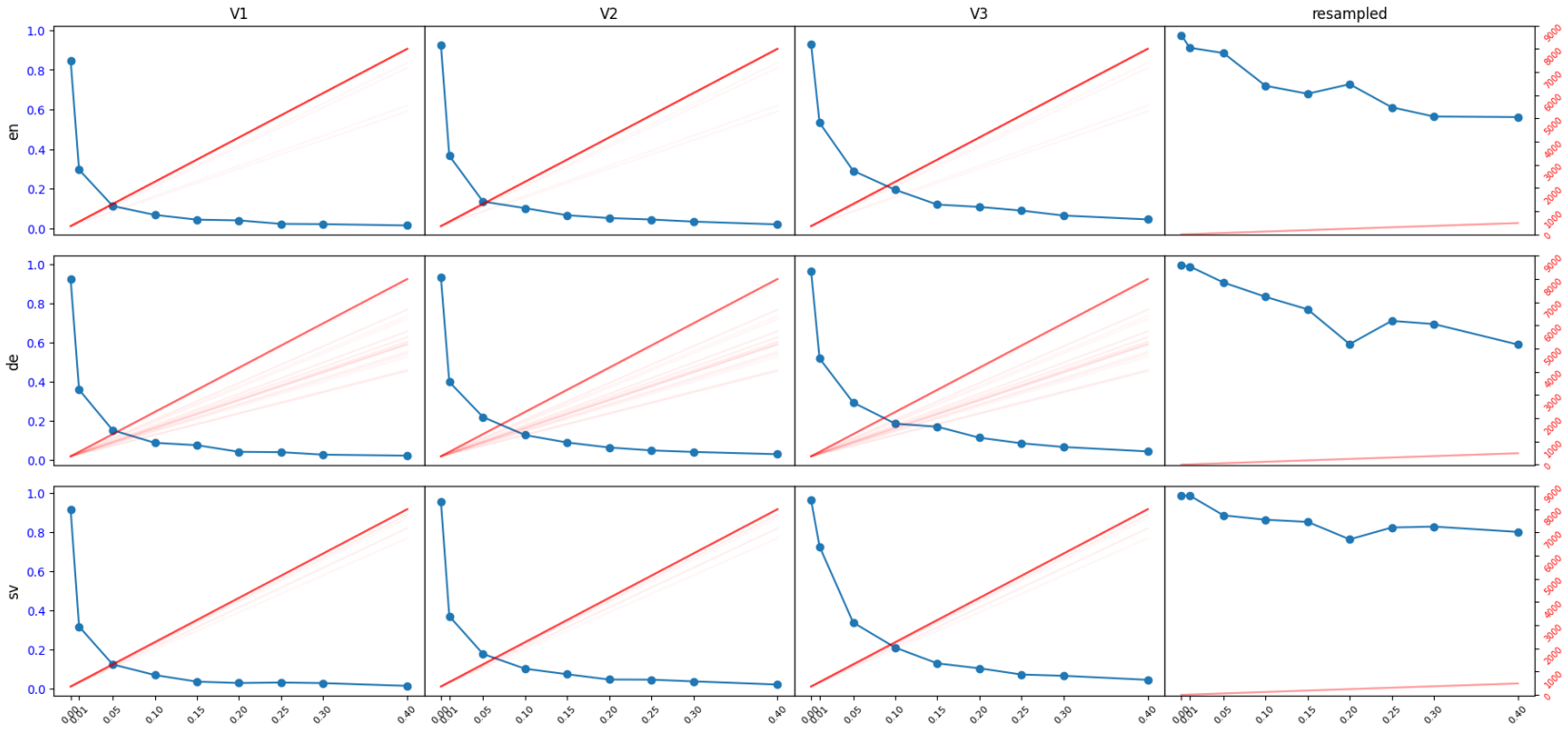


Figure 8: ARI of DWUG DE/EN/SV clusters over increasing percentages of noisy edges. The right y-axis (in *red*) shows the raw number of noisy edges. The x-axis shows the percentage of perturbed edges.

Replicability of word sense distributions

	min	avg	max
DE 1–4	.0	**.10**	.28
DE 1–5	.0	**.08**	.20
EN 1–4	.11	**.22**	.45
EN 1–5	.0	**.19**	.42
SV 1–4	.0	**.19**	.48
SV 1–5	.0	**.10**	.42

Table 4: JSD between sense distributions for DWUG DE/EN/SV rounds 1–4 and 1–5 compared to `resampled` datasets.

Conclusion

- we added **thousands of judgments** to existing WUG datasets making them more **densely annotated** and **reliable**
- we found that
  - clustering **quality increases** with annotation rounds
  - original datasets were **not optimal**, results should be reconsidered
  - final clusterings have **high validity**
  - clusterings derived on sparsely annotated graphs are **prone to annotation noise**
  - word sense distributions can often be approximated well with **smaller samples** and **random edge sampling**
- **main conclusion**: large samples of uses should be sacrificed in favor of **large samples of edges**
- datasets can be used to tune and evaluate models for a multitude of tasks, such as **WiC, WSI** and **LSCD**[2]

[2] Find the datasets at `www.ims.uni-stuttgart.de/data/wugs`

Future work

- Will the **improved data quality** lead to **higher performance** of WSI and LSCD models?
- Can previous results on performance relations be **reproduced** with the more reliable data?
- Can we improve the clustering quality through **alternative clustering algorithms**?
- Can we find **efficient** and **robust** node and **edge sampling** strategies?
- What are **alternative ways of evaluating** the quality of the annotation, the clustering or the change scores?

References I

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

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/

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., 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