# More DWUGs

## Extending and Evaluating Word Usage Graph Datasets in Multiple Languages

**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

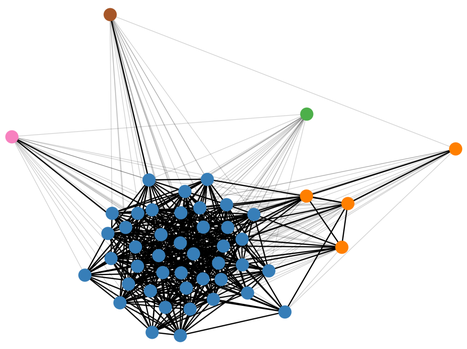
## Introduction

- **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

## 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.

## A WUG example



## Annotation

- **DWUG** [Schlechtweg et al. 2021]
  - English, German, Swedish
  - widely used
  - **many uses** per word (≤ 200)
  - sophisticated edge sampling
  - 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

## Statistics

| | $\lvert E\rvert$ | | | $+\lvert J\rvert$ | | | $\lvert J\rvert$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | DE | EN | SV | DE | EN | SV | DE | EN | SV |
| 1–4 | .03 | .02 | .02 | 74 | 69 | 48 | 40K | 36K | 24K |
| 1–5 | .03 | .03 | .03 | 142 | 193 | 191 | 48K | 46K | 37K |
| **1–6** | **.05** | **.05** | **.05** | **297** | **487** | **394** | **63K** | **69K** | **55K** |
| resampled | .45 | .35 | .60 | | | | 10K | 7K | 16K |

Table: Coverage for DWUG datasets. $\lvert E\rvert$: avg. % of annotated edges, $+\lvert J\rvert$: avg. increase in number of judgments, $\lvert J\rvert$: absolute number of judgments.

## Validity

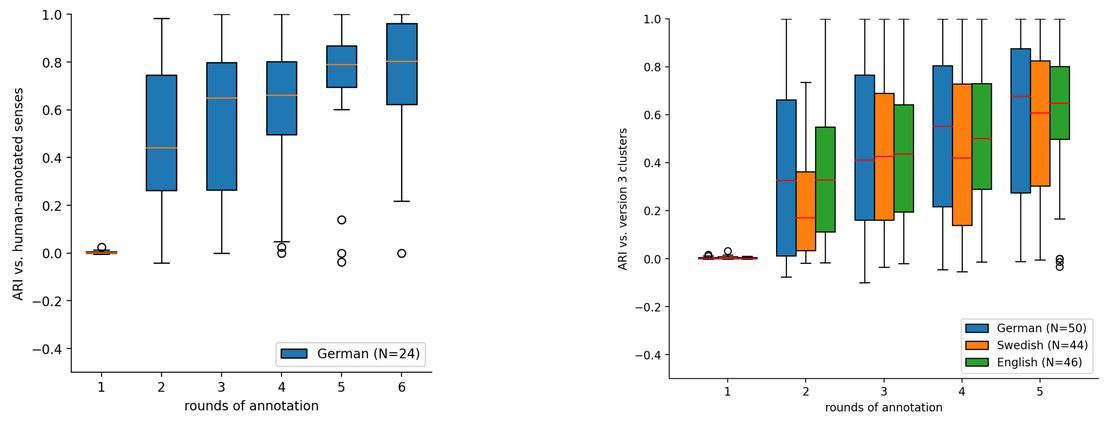



Figure: Left: ARI of DWUG DE clusters by round vs. DWUG DE Sense annotation. Right: ARI of DWUG DE/EN/SV clusters by round vs. round 6 clusters.

## Robustness

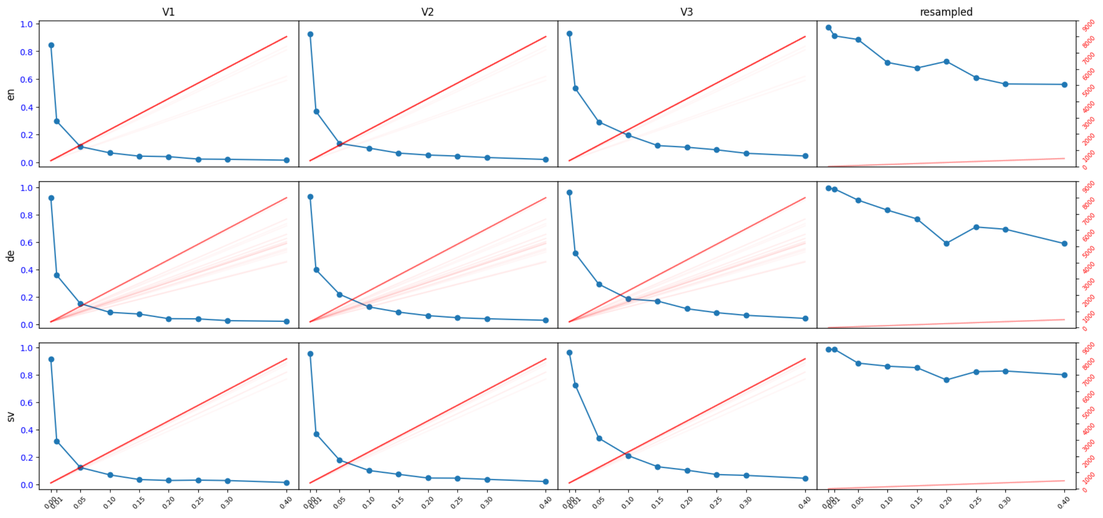



Figure: ARI of DWUG DE/EN/SV clusters over increasing percentages of noisy edges. The right y-axis (in *red*) shows the raw number of noisy edges.

## 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**:

`www.ims.uni-stuttgart.de/data/wugs`

## Replicability

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

References

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`