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

Abstract

Word Usage Graphs (WUGs) represent human semantic proximity judgments for pairs of word uses in a weighted graph, which can be clustered to infer word sense clusters from simple pairwise word use judgments, avoiding the need for word sense definitions. SemEval-2020 Task 1 provided the first and to date largest manually annotated, diachronic WUG dataset. In this paper, we check the robustness and correctness of the annotations by continuing the SemEval annotation algorithm for two more rounds and comparing against an established annotation paradigm. Further, we test the reproducibility by resampling a new, smaller set of word uses from the SemEval source corpora and annotating them. Our work contributes to a better understanding of the problems and opportunities of the WUG annotation paradigm and points to future improvements.

1 Introduction

In recent years, a new annotation paradigm for word senses has emerged under the name of **Word Usage Graphs** (WUGs, Schlechtweg et al., 2020, 2021d). In this paradigm, humans provide semantic proximity judgments of pairs of word uses (also known as *Word-in-Context*, WiC), which are then represented in a weighted graph and clustered with a graph clustering algorithm, as displayed in Figure 1 showing a selection of clustered graphs from multiple WUG datasets. In this way, word sense clusters can be inferred from simple pairwise word use judgments, avoiding the need for a sense inventory, which can be tedious to create. While, up to now, this approach has been applied mainly within the field of Lexical Semantic Change Detection (LSCD) (e.g. Kurtyigit et al., 2021; Zamora-Reina et al., 2022; Chen et al., 2023), it can be applied generally in a Word Sense Induction (WSI) setting (Aksenova et al., 2022) or for Word Sense Disambiguation (WSD) when combined with a sense labeling procedure for word sense clusters (cf. Giulianelli et al., 2023).

As a recent approach to the study of word senses, the WUG annotation paradigm brings many open questions and uncertainties relating to reduction of annotation load (see Section 4.1), clustering of the annotated graphs as well as the stability and reproducibility of the resulting clusters. In this paper, we try to answer some of these questions relying on the first large WUG dataset created in the SemEval-2020 Task 1 on Unsupervised Lexical Semantic Change Detection (Schlechtweg et al., 2020). We add additional rounds of annotation to the English, German and Swedish datasets, to more densely populate the graphs and make the inferred sense clusters and semantic change scores more reliable. This allows a comparison against the earlier inferred clusters and scores. We also compare the inferred clusters over annotation rounds against an external gold standard of sense definition annotations in German (Schlechtweg et al., 2024c). In this way, we test the **validity** of previous clusterings. We also evaluate cluster **robustness** by measuring the clustering variance with different degrees of noise introduced to the annotation. Further, to test the **replicability** of the SemEval data, we completely resample and annotate a small new set of word uses from the SemEval source corpora. We hope this paper contributes to a better understanding of the problems and opportunities of the WUG annotation paradigm and points to future improvements in the process. We summarize our contributions going beyond previous work: (i) Our work adds thousands of additional judgments to the datasets created by Schlechtweg et al. (2021d) and Kurtyigit et al. (2021) and thus makes them much more densely annotated.[1] These datasets can be

[1] The updated datasets can be found at www.ims.uni-stuttgart.de/data/wugs. The updated datasets are DWUG EN/DE/SV V3.0.0, DWUG EN/DE/SV resampled V1.0.0 and DiscoWUG V2.0.0.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 14379–14393
November 12-16, 2024 ©2024 Association for Computational Linguistics

used to tune and evaluate models for a multitude of tasks, such as WiC, WSI and LSCD (Schlechtweg et al., 2024a). (ii) We resample and annotate data from the same sources as Schlechtweg et al. (2021d) providing an independent replication of their results (word sense clusters, change labels), which can serve as a comparison and a way to evaluate reliability and accuracy of the original dataset. (iii) We are the first to provide a solid way to validate the cluster derivation approach defined by Schlechtweg et al.: We compare the obtained clusters to clusters obtained by an independent (more traditional) annotation approach. Additionally, we propose to evaluate the clusters from earlier rounds to the clusters obtained on the full data. Both approaches show that the original data was not optimal, quality improves over rounds and that the final datasets created by us have near to optimal quality.

The paper is structured as follows: Next, we introduce related work on word sense annotation approaches. Then, we introduce the datasets we will extend or replicate followed by a description of our annotation procedure. We then describe a number of experiments on the resulting datasets testing validity, robustness and replicability of the derived sense clusters. This is followed by concluding remarks and an outlook to future work.

2 Related Work

Existing word sense annotation procedures can be distinguished into three main categories: (i) use-sense, (ii) lexical substitution and (iii) use-use annotation (cf. Erk et al., 2013). (i), use-sense annotation, has a long tradition within the task of WSD (Weaver, 1949/1955). Annotators usually choose the best-fitting word sense definition for a word use as in this example:

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

In annotation of lexical substitutes, annotators are presented only with a single word use and asked to provide other words which could be substituted for the target word (McCarthy and Navigli, 2009). Consider this example:

> **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.

Possible substitutes for *arms* in this use would be *weapons* or *guns*. Although it is not a widespread approach, it is possible to represent such lexical substitutes as vectors, to derive similarities from these and to represent them in a weighted graph which can be clustered (McCarthy et al., 2016).

In use-use annotation, typically pairs of uses are judged according to their semantic proximity (i.e., relatedness or similarity):

> **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.

Related approaches obtain pairwise judgments using a comparative annotation framework (Abdalla et al., 2023), or infer them through spatial arrangement (Majewska et al., 2021), sentence sorting (Ramsey, 2022), use-sense judgments (Pilehvar and Camacho-Collados, 2019) or difficulty estimation (Alfter et al., 2022). Computational modeling of use-use proximity judgments on binary or graded scales has seen a recent upsurge under the label "Word-in-Context task" (Pilehvar and Camacho-Collados, 2019; Armendariz et al., 2020; Cassotti et al., 2023b).

The WUG approach builds on use-use proximity judgments by exploiting their transitive interconnectedness through graph representation (McCarthy et al., 2016; Schlechtweg et al., 2021d). There are a number of recent WUG datasets for multiple languages (Schlechtweg et al., 2021d; Kurtyigit et al., 2021; Baldissin et al., 2022; Zamora-Reina et al., 2022; Kutuzov et al., 2022; Aksenova et al., 2022; Chen et al., 2023), most of them with a diachronic component by sampling uses from different time periods. A few studies investigate the clustering and/or edge sampling procedures (Schlechtweg et al., 2021a; Tunc, 2021; Kotchourko, 2021). See also Schlechtweg (2023, pp. 54–67) for an in-depth analysis of cluster errors and the robustness of clusterings and change scores derived from them. Most related to our work are the WUG datasets created for SemEval-2020 Task 1 (Schlechtweg et al., 2020, 2021d) as these are used and extended for this study.

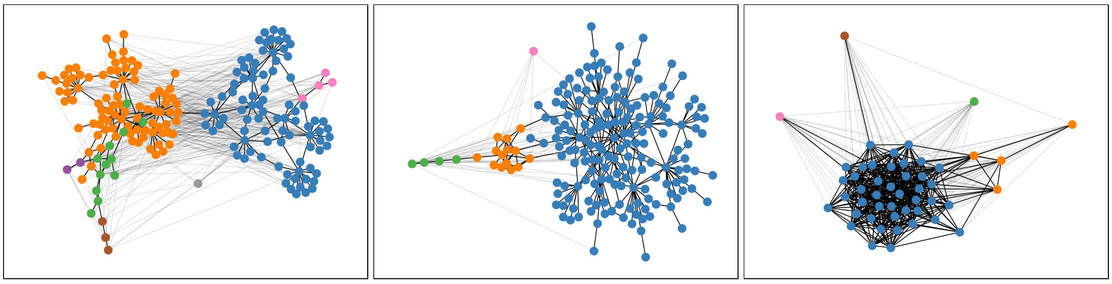

Figure 1: WUGs from DWUG V1 (SemEval clustering) and DiscoWUG V1: English *plane* (left), Swedish *färg* (middle) and German *anpflanzen* (right). Isolates were removed.

3 Word Usage Graphs

A WUG $\mathbf{G} = (\mathbf{U}, \mathbf{E}, \mathbf{W})$ is a weighted, undirected graph, where nodes $u \in U$ represent word uses and weights $w \in W$ represent the semantic proximity of a pair of uses (an edge), $(u_1, u_2) \in E$ (McCarthy et al., 2016; Schlechtweg et al., 2020). The set of uses U can be sampled from different time periods t_1, t_2 ... t_n, where we refer to the time-specific use subsets as U_1, U_2 ... U_n. In practice, semantic proximity can be measured by human annotator judgments on a scale of relatedness (Brown, 2008; Schlechtweg et al., 2018) or similarity (Erk et al., 2013). WUGs obtained from the annotation are often sparsely observed and noisy. This poses a very specific problem that calls for a robust clustering algorithm. Hence, Schlechtweg et al. (2021d) implement a variation of correlation clustering (Bansal et al., 2004) which minimizes the sum of cluster disagreements, i.e., the sum of low edge weights (semantic proximity) within a cluster plus the sum of high edge weights across clusters. For this, one has to choose a threshold h on edge weights deciding which weights will be considered as high and which ones as low. Schlechtweg et al. set $h = 2.5$. Consequently, the weight $W(e)$ of each edge $e \in E$ in a WUG $\mathbf{G} = (\mathbf{U}, \mathbf{E}, \mathbf{W})$ is shifted to $W'(e) = W(e) - 2.5$ (e.g. a weight of 4 becomes 1.5). Those edges e with a weight $W'(e) \geq 0$ are referred to as **positive** edges P_E while edges with weights $W'(e) < 0$ are called **negative** edges N_E. Let further $C : U \mapsto L$ be some clustering on U, $\phi_{E,C}$ be the set of positive (high) edges **across** any of the clusters in clustering C and $\psi_{E,C}$ the set of negative (low) edges **within** any of the clusters. The algorithm then searches for a clustering C that minimizes the sum of weighted cluster disagreements:

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

That is, the sum of positive edge weights between clusters and (absolute) negative edge weights within clusters is minimized. Minimizing SWD is a discrete optimization problem which is NP-hard (Bansal et al., 2004). As most WUGs have a relatively low number of nodes (≤ 200), Schlechtweg et al. chose to approximate the global optimum with Simulated Annealing (Pincus, 1970), a standard discrete optimization algorithm. In order to reduce the search space, the algorithm iterates over different values for the maximum number of clusters (≤ 20). It also iterates over randomly as well as heuristically chosen initial clustering states.[2]

The finally obtained clustering $C : U \mapsto L$ maps each use $u \in U$ to a cluster label $l \in L \subset \mathbb{N}$. From this, Schlechtweg et al. calculate a **cluster (sense) frequency distribution** D encoding the size of each cluster as

$$D = (f(L_1), f(L_2), ..., f(L_i))$$

where $L_i < L_{i+1}$ and $f(L_i)$ is the number of times any use from U was mapped to the cluster label L_i (cf. McCarthy et al., 2004; Lau et al., 2014). Correspondingly, they obtain two distributions D_1, D_2 from C for the two time-specific use sets U_1, U_2. D, D_1 and D_2 are ordered and contain the frequencies for the full set of cluster labels L so that the ith index always corresponds to the same cluster label. (Note that this means that the time-specific sense frequency distributions are obtained from clustering the full graph.) From the time-specific D_1 and D_2, Schlechtweg et al. calculate a binary and a graded change score, respectively as cluster gain or loss (binary) and the Jensen-Shannon distance between D_1 and D_2 (graded) (Lin, 1991; Donoso and Sanchez, 2017).

[2] Find their code at: https://github.com/Garrafao/WUGs.

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

Table 1: The DURel relatedness scale (Schlechtweg et al., 2018).

4 Data

We select the DWUG dataset (Schlechtweg et al., 2021d) because it was richly annotated in multiple rounds of annotation and has been widely used (e.g. Cassotti et al., 2023a; Giulianelli et al., 2023; Kutuzov et al., 2024). The data contains use pairs of English, German and Swedish target words annotated on the scale in Table 1. For each word, the annotations were represented in a weighted (use-use) graph by taking the uses as nodes, use pairs as edges and the median relatedness judgment as edge weights. As described in Section 3, use clusters were computed on each graph using the Correlation Clustering algorithm and from the clustering binary and graded change scores for each word were inferred, reflecting changes in clusters over time. Find clustered WUG examples in Figure 1.

Additionally, we use the DiscoWUG dataset (Kurtyigit et al., 2021) as it extends the DWUG DE dataset by a number of target words with uses sampled from the same time-specific corpora as DWUG. For evaluation of the annotation quality, we use DWUG DE Sense (Schlechtweg et al., 2024c) which annotates a subset of uses sampled from the German DWUG dataset with traditional use-sense annotations.[3] Find an overview of the datasets in Table 5 in Appendix A.

4.1 DWUG DE/EN/SV

The DWUG dataset of Schlechtweg et al. (2021d), which we refer to here as DWUG V1, was created over four rounds of annotation. (See Table 2 for an overview of all dataset versions.) Each round of annotation used *combination* and *exploration* sampling criteria based on the WUG resulting from the previous rounds of annotation. These criteria had the goal of producing graphs with a faithful clustering (i.e., one that approximates the clustering that would be obtained on a fully-annotated graph) while reducing the total number of judgments necessary. The annotation of the full graph for each word is infeasible due to the quadratic number of available edges. Hence, the resulting graphs are incomplete (sparsely observed), see graphs in Figure 1. This issue was exacerbated by the rather large number of uses per word contained in the dataset (see Table 5 in Appendix A). Also, the annotation procedure had to be stopped after round 4, before the convergence criterion (all inferred clusters connected by at least one edge) was met. Hence, some graphs have unconnected clusters, see *färg* in Figure 1.

4.2 DiscoWUG

The DiscoWUG dataset of Kurtyigit et al. (2021), which we refer to here as DiscoWUG V1, was created in one round of annotation. Uses were sampled from the same corpora as the DWUG DE data. It can thus be seen as an extension of DWUG DE in terms of target words. However, much less uses (25+25) were sampled per target word for DiscoWUG. Hence, the graphs are more densely annotated and suffer less from unconnected clusters (see Table 2). Target words were selected partly from the top-predicted graded change scores of computational models and partly at random. This inhibits the datasets applicability in model evaluation as models run the risk of being evaluated in a circular process.[4] Another difference to DWUG DE is the sampling strategy for edges: for DiscoWUG edges were sampled purely randomly. We see this as a beneficial property of the graphs as it allows less biased statistical inference.

4.3 DWUG DE Sense

The DWUG DE Sense dataset reannotates a subsample from DWUG DE target words and uses with binary use-sense judgments (Schlechtweg et al., 2024c). This allows to compare sense-related statistics inferred on DWUG DE to be evaluated against an independent operationalization strategy (pp. Schlechtweg, 2023, 58–59). 24 target words (out of 50) were randomly chosen from the DWUG DE dataset together with extracted sense definitions from two historical dictionaries (Paul, 2002; DWDS, 2021). Then 50 randomly sampled uses for each target word (25 per time period from at most 100) were annotated. Each use was annotated by three annotators with the sense definition best describing the meaning of the target word in

[3] All pre-existing datasets can be accessed at: www.ims.uni-stuttgart.de/data/wugs

[4] Kurtyigit et al. (2021, Table 3) observe that model correlations on predicted target words are low.

		AN	|J|	|E|	UNC	AV	SPR	KRI	Sample	R	Reference
DWUG DE	1.1.0	8	817.66	2.75	0.92	1.81	0.60	0.66	SemEval	1–4	(Schlechtweg et al., 2021c)
	2.3.0	8	959.58	3.39	0.34	1.73	0.61	0.67	SemEval	1–5	(Schlechtweg et al., 2022b)
	3.0.0	11	1256.74	4.90	0.58	1.53	0.61	0.67	SemEval	1–6	this paper
DWUG EN	1.0.0	8	816.30	2.48	2.02	1.72	0.53	0.61	SemEval	1–4	(Schlechtweg et al., 2021b)
	2.0.1	9	1009.07	3.18	0.87	1.66	0.56	0.63	SemEval	1–5	(Schlechtweg et al., 2022a)
	3.0.0	13	1495.76	5.09	0.85	1.49	0.55	0.63	SemEval	1–6	this paper
DWUG SV	1.0.0	5	660.61	2.15	0.86	1.56	0.62	0.68	SemEval	1–4	(Tahmasebi et al., 2021)
	2.0.0	7	851.80	2.73	0.68	1.59	0.63	0.67	SemEval	1–5	this paper
	3.0.0	13	1245.64	4.61	2.05	1.38	0.63	0.67	SemEval	1–6	this paper
DiscoWUG	1.1.1	8	341.75	25.25	0.07	1.09	0.64	0.58	random	1	(Kurtyigit et al., 2022)
	2.0.0	11	360.91	24.82	0.00	1.18	0.62	0.57	unc.	2	this paper
resampled DE		3	670.67	44.75	0.00	1.21	0.70	0.59	SemEval	1	this paper
resampled EN		3	490.40	35.29	0.00	1.13	0.59	0.56	SemEval	1	this paper
resampled SV		6	1064.47	59.85	0.00	1.40	0.65	0.56	SemEval	1	this paper

Table 2: Overview of data statistics. Version 1 datasets include rounds 1–4 of annotation, version 2 includes rounds 1–5, and Version 3 rounds 1–6. AN = number of annotators, $|J|$ = average of number of judged usage pairs per word, $|E|$ = avg. percentage of edges annotated, UNC = avg. no. of uncompared multi-cluster combinations, AV = avg. no. of judgments per usage pair, SPR = weighted mean of pairwise Spearman, KRI = Krippendorff's alpha, Sample = sampling strategy, R = annotation round.

this use. The annotations were cleaned and aggregated in different conditions. We use the 'maj$_3$' aggregation for our experiments where all uses are guaranteed to have perfect agreement from the three annotators.

5 Annotation

We recruited at least three annotators per language. Most were current students or had university-level education. All of them were native speakers of the respective language.[5] They were instructed in a short training using a version of the DURel guidelines (Schlechtweg et al., 2024b). Below, we first describe for each dataset how uses and edges were sampled for annotation. Then, we describe how the annotated data was postprocessed and analyzed.

5.1 DWUG DE/EN/SV additional rounds

Because the annotation process for V1 of the DWUG datasets was stopped early, we continue the annotation algorithm with two more rounds of annotation. Round 5 was sampled with the same criteria as rounds 2–4. Round 6 was sampled with a similar, but simplified process using only two different sampling heuristics—the `random` heuristic, which samples use pairs uniformly at random from the pool of edges; and the `unconnected` heuristic, which samples edges from pairs of clusters that had not yet been compared explicitly. Round 5 was annotated in the same way as rounds 1–4 with the use of spreadsheets provided to annotators. Round 6 was annotated using the DURel annotation tool (Schlechtweg et al., 2024b).[6] For 6`[unc]` we sampled at most 3 edges (use pairs) per unconnected cluster. All annotators annotated the same edges in the same order. For 6`[rnd]`, annotators were instructed to annotate a random sequence of use pairs presented to them by the annotation system for 30 minutes. The sequence was different for each annotator. For Swedish, we excluded words which were not used for the SemEval task (mostly due to a high number of *cannot decide* judgments and low agreement). Because of low agreement for Swedish in the first try, the study was repeated with new annotators. This time annotators had 60 minutes for the 6`[rnd]` portion. The Swedish data from the repetition was then added to the first-try data for cleaning and aggregation described below.

For all languages, the dataset that considers rounds 1–5 of annotation is called DWUG V2; that is, DWUG V2 is DWUG V1 plus one additional round of annotation. Likewise, DWUG V3 considers rounds 1–6. See Table 2 and Figure 4 in the appendix for an overview of each dataset with corresponding rounds.

5.2 DiscoWUG

We annotated the latest version (1.1.1) of the DiscoWUG dataset using the `unconnected` heuristic. DiscoWUG contains much less uses per word than the other datasets (50 vs. up to 200) and

[5] All annotators were paid according to the standard in the country in which they were employed.

[6] www.ims.uni-stuttgart.de/data/durel-tool

	|E|			+|J|		
	DE	EN	SV	DE	EN	SV
1	0.28	0.27	0.28	-	-	-
1–2	1.92	1.69	1.47	489	504	354
1–3	2.42	2.19	1.96	171	161	159
1–4	2.75	2.48	2.15	74	69	48
1–5	3.39	3.18	2.73	142	193	191
1–6	4.90	5.09	4.61	297	487	394
resampled	44.75	35.29	59.85	-	-	-

Table 3: Coverage by round. $|E|$ = average percentage of edges annotated in the combined datasets, $+|J|$ = average increase in number of judgments per word from the previous round. Both statistics only include lemmas that were considered in all rounds of annotation.

hence much less unconnected clusters and suffers much less from sparsity in general, see column 'UNC' in Table 2. Hence, we deemed it sufficient to merely connect the few unconnected clusters without using the random heuristic. The data was annotated in the DURel annotation tool.

5.3 Resampled

For this annotation portion, we chose 15 words randomly from each DWUG dataset. For each word, we sampled 25 uses per SemEval time period, resulting in 50 uses.[7] These uses were then uploaded to the DURel tool and annotators were instructed to annotate the random sequence of use pairs displayed to them by the system, for each word. They were instructed to spend 60 minutes on each word. Because of the lower number of uses than in the original DWUG data, the resulting graphs are more densely connected, see column 'UNC' in Table 2. Similar to the DWUG round 6 study described in Section 5.1, for Swedish non-SemEval target words were excluded and the study had to be repeated once. With the data from this study we aim to replicate results from the DWUG dataset.

5.4 Postprocessing and analysis

The data from the annotation rounds presented in this paper was post-processed similarly to previous rounds. Judgments where annotators indicated *cannot decide* were removed from the data for agreement calculation. We de-duplicated multiple judgments by the same annotator. In the small handful of cases where there was no self-agreement, these judgments were removed. To assess the reliability of annotations on an annotator level, we computed pairwise agreement between annotators, including judgments from previous rounds of annotation, where there was overlap. Then, for each annotator we took an average of the pairwise agreement scores with other annotators, weighted by the number of items they overlapped on. Two annotators from the English data (round 6 and resampled) and three from the Swedish data (one from round 6 and two from both round 6 and resampled) were excluded due to having relatively low mean pairwise agreement.[8] From the cleaned data, we computed WUGs, clusters and change scores for different data versions (see Table 2) by the procedure described in Section 3. We use the WUG pipeline with default opt parameters to generate graphs, cluster them and compute statistics and change scores.

Find important statistics for major versions (rounds of annotation) of each dataset in Table 2. Versions accumulate data from all previous rounds (see column 'R'). The clusterings for each version were hence obtained on the full data from previous rounds. In the experiments reported in Section 6, we use the dataset versions given in Table 2 and split DWUG into annotation rounds as shown in Table 3.

In summary, Table 2 shows that subsequent versions of the data not only include more annotations, but cover a larger portion of the total graph. This is especially true for the V3 datasets introduced in this paper. The average number of uncompared clusters is below 2 in all studies, though sometimes slightly higher with more data. This may be due to the discovery of additional clusters as the number of judgments increases. Agreement remains strong for all versions of data.

6 Experiments

We now evaluate the validity of the inferred clusters over rounds of annotation. This is followed by tests of the robustness of the final clusterings, and their replicability through a complete resampling and reannotation of data.

6.1 Validity of clusters

First, we assess what progressive rounds of annotation contribute to the quality of the resulting WUGs. Naturally, the proportion of the complete

[7] Note that we did not apply the SemEval constraint on sentence length when sampling uses.

[8] Mean pairwise Krippendorff's α was -0.05 and 0.46 for the excluded English annotators and -0.24, 0.21, and 0.26 for the excluded Swedish annotators.

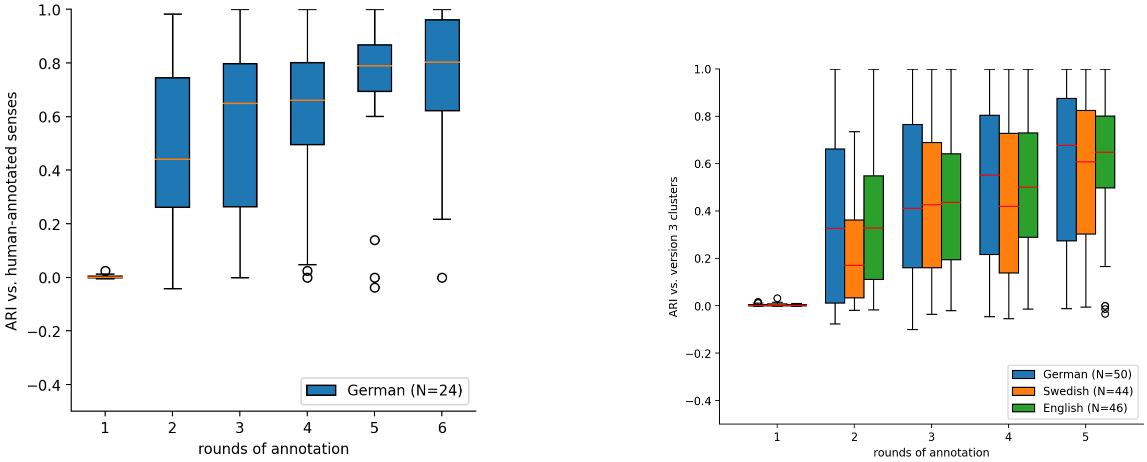

Figure 2: Left: ARI of DWUG DE clusters after each round vs. DWUG DE Sense annotation. Spreads indicate variation over lemmas (N=19); only lemmas appearing in all WUG datasets and the sense annotation dataset are included. Right: ARI of DWUG DE/EN/SV clusters vs. V3. Spreads indicate variation over lemmas. Only lemmas that were annotated in each round are included. Uses that were assigned to the noise cluster round 6 were excluded from the ARI computation.

graph that was annotated rises with each successive round of annotation. That said, only a very small fraction of the edges of the total graph were annotated, even considering all six rounds.

Adding successive rounds of annotation, we clustered the resulting WUGs with Correlation Clustering (Section 5.4) where the resulting clusters ideally correspond to different senses of the word. We performed two experiments to test the quality of the WUG clusters after each round of annotation. Additionally, in Appendix B, we measure the accuracy of the semantic change scores/labels derived from the clusters (see Section 3) to more directly estimate the reliability of the datasets on the LSCD level.

External gold standard For German, we compared the cluster assignments of uses to traditional word sense annotations present in DWUG DE Sense, see Figure 2, left. Cluster quality is assessed using the Adjusted Rand Index (ARI; see Fahad et al., 2014), which is defined as follows:

$$ARI = \frac{RI - Expected_{RI}}{max(RI) - Expeted_{RI}}$$

Here, RI stands for the Rand Index, which measures the number of pair agreements within the data — that is, pairs of instances (in our case ‘uses’) that are correctly placed in the same or different clusters. The $Expetcted_{RI}$ is the expected number of such agreements by chance, calculated based on the distribution of the clusters, while the $max(RI)$ is the maximum possible value of RI, which occurs when all pairs are classified perfectly.

We see that the first five rounds of annotation have a clear positive effect on the quality of the clusters. The median ARI increases, and just as importantly, word-level variance goes down. This suggests that the amount of data needed for cluster quality depends a lot on the word itself, an issue we explore further in Section 6.2. Moreover, the WUGs that include all six rounds of annotation (corresponding to dataset V3) are very well correlated with the sense annotations, achieving a median ARI of 0.81. We note, however, that the sixth round of annotation shows only minimal improvement in median ARI over the fifth.

Final clustering Next, we compare the above-described clusterings obtained on successive rounds of annotation to the final clustering of the WUG constructed with all six rounds of annotation (Figure 2, right). The main motivation for this experiment is to compare the cluster quality also for of the other two languages where explicit sense annotation does not exist. Correspondence to the clustering of the final WUG increases as successive rounds of annotation are added, and this is observed for all languages. Assuming (as is suggested by Figure 2, left), that the final round has the best quality clustering, this demonstrates that collecting additional rounds of annotation with new annotators was beneficial for the quality of clusters across all three languages, at least up to the amount of data represented by rounds 1–5.

6.2 Robustness of clusters

In this section, we assess the robustness of the clustering method to errors within the annotation process. In particular, we conduct an experiment, similar to Schlechtweg et al. (2021d) to assess the resilience of our results when exposed to (non-meaningful) variations in graph weights. In this experiment, randomly generated annotations are added and the edge weights are recomputed taking into account both the original and the noisy annotations. We then generate new graphs and perform clustering on these modified graphs. We carry out a comparative analysis of the correspondence between the clusters in the original graphs, i.e., our ground truth, and those in the manipulated graphs. As in Section 6.1, cluster correspondence is measured with ARI.

This investigation, which includes V1, V2, V3 and the `resampled` datasets as shown in Figure 3, reveals a key finding: the stability of the cluster structure in the graphs is significantly influenced by the number of edges. We first assess the cluster variance and recalculate the clustering a second time, i.e., adding 0% noisy annotations. Results shows high ARI scores, ranging 0.84-0.98.[9] A drop in performance is immediately noticed when introducing 0.01% of random annotations. However, it is important to highlight that this represent the harshest scenario, which assumes a completely random annotator. Increasing the number of annotated edges (from V1 to V2 and V3) leads to more robust results. In fact, for the `resampled` dataset, which has a much higher proportion of annotated edges, the ARI scores for English, German and Swedish remain above 50% even after introducing 40% of noisy edges.

While it is necessary to generate a sufficient number of uses (nodes) for the annotation sample to be sufficiently representative of the underlying corpus, the robustness analysis demonstrates that the number of edges (of pairs of annotated uses) also plays a fundamental role for data quality. However, there is a tension between the need for a representative sample and the challenges of annotating a sufficiently large number of edges because of the quadratic relationship between the number of nodes and the corresponding number of edges in a fully connected graph. One way to address this issue could involve focusing more annotation efforts on words with a higher degree of polysemy, while reducing effort for monosemous words (Appendix C).

	min	avg	max
DE V1	.0	.10	.28
DE V2	.0	.08	.20
EN V1	.11	.22	.45
EN V2	.0	.19	.42
SV V1	.0	.19	.48
SV V2	.0	.10	.42

Table 4: Jensen-Shannon distance between sense distributions for V1 and V2 compared to `resampled`.

We provide a similar robustness evaluation for change scores in Appendix B.

6.3 Replicability of clusters

We now evaluate DWUG `resampled` data described in Section 5.3 against V1 and V2 in order to understand how well we can replicate the SemEval (and subsequent) annotation efforts with a small sample of uses annotated densely with a simple edge sampling approach (random). The `resampled` datasets contains a subset of 15 words that are also found amongst the respective V1 and V2 words, but word uses were resampled from the source corpora and thus have a lesser overlap with V1 and V2.[10] The data was annotated much more densely (cf. Table 3), resulting in a much denser graph; we surmise that a denser graph will be clustered more reliably than a sparse graph. However, through sampling variability the use samples could be less representative in `resampled` than in V1 and V2.

For each word in `resampled`, we calculate the Jensen-Shannon distance (JSD) between the sense distributions in `resampled` and in the respective version from DWUG. Table 4 shows the average value over all words.[11] As the number of sense clusters may be different between datasets, we address this issue by trimming the longer distribution to match the shorter one.

Overall, we can observe that the distance in sense distributions is rather low for all datasets (between 0.08 and 0.22). The maximum distance observed is 0.48, but for some data the maximum distance is much lower, 0.2–0.28. The zero val-

[9] The lower ARI score (0.84) is obtained for English V1 where we observe a high number of randomly clustered nodes (Kotchourko, 2021), i.e., nodes with all incoming edges equal to 2.5.

[10] The overlap between *uses* is 1% for English, 3% for Swedish, 4% for German.

[11] For comparability with previous results, we set the base of the logarithm to 2, which results in a number (0–1).

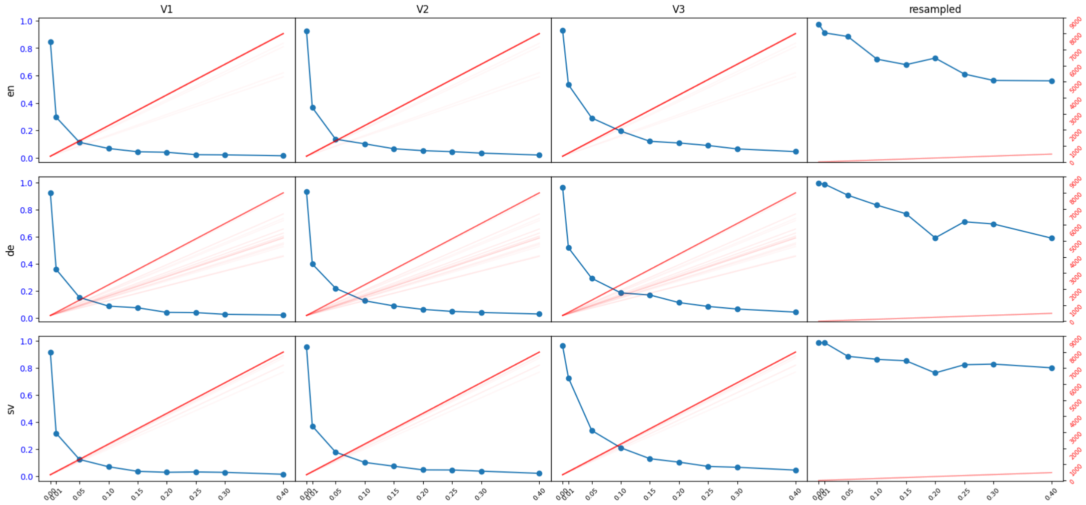

Figure 3: Robustness - ARI scores computed with respect to increasing percentages of noisy edges. The right y-axis (in *red*) shows the raw number of noisy edges. The x-axis shows the percentage of perturbed edges.

ues show that some words are perfectly aligned, thus increasing the uses up to 4 times per time period (as in DWUG) for these words does not add more information. The German data has generally the lowest average and maximum distances. This suggests that it is possible to approximate the very costly SemEval annotation result with a simple, less costly annotation procedure. For all languages, the distance to the `resampled` data becomes lower in V2. As we know from Section 6.1, V2 also improves in cluster quality, suggesting that some of the distance is explained by lack of quality in the SemEval data.

7 Conclusion

In this paper, we tested the validity, robustness and replicability of the largest existing WUG datasets used in SemEval-2020 Task 1. We added thousands of additional judgments to the datasets making them much more densely annotated and reliable. These datasets can be used to tune and evaluate models for a multitude of tasks, such as WiC, WSI and LSCD. Then, we reclustered the data based on increasing amounts of annotation and found that clustering quality increases with annotation rounds. This also shows that the original SemEval datasets were not optimal for evaluation, and possibly results should be reconsidered. Our robustness analysis supports this finding and suggests that small sample sizes of uses, leading to more densely annotated graphs as the proportion of annotated edges is higher, have a considerable effect on robustness of clusterings. We further resampled and reannotated the data providing an independent replication showing that the SemEval word sense distributions can often be approximated well with smaller samples and simpler (random) edge sampling. The main conclusion from our work is that in future annotation studies large samples of uses should be sacrificed in favor of large samples of edges, in order to create more densely annotated graphs. This aligns with and strengthens the findings of Kutuzov and Pivovarova (2021) and Zamora-Reina et al. (2022).

A natural hypothesis for future work is whether the improved data quality will lead to higher performance of WSI and LSCD models and whether previous results on performance relations can be reproduced with the more reliable data. Further interesting questions concern the improvement of the annotation procedure: Can we improve the clustering quality? 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?

8 Limitations

Although we tried to equalize conditions across annotation rounds, some factors differ: In rounds 1–5, judgments were provided in simple spreadsheets while for round 6 we used the DURel annotation tool. This led to a difference how use pairs were presented to annotators: In the spreadsheets pairs were randomized across lemmas while in the DURel tool annotators typically judge one lemma at a time. The tool also suffered from minor bugs during the time of annotation which may have had an influence on the data. Furthermore, uses for the resampled study were not sampled from the source corpora in the exact same way as in SemEval as for the latter we applied a constraint on minimum sentence length which was not applied for the former.

Acknowledgments

DS, PC, BN, DA and NT have been funded by the research program 'Change is Key!' supported by Riksbankens Jubileumsfond (under reference number M21-0021). The work has also been supported by SemRel Group (DFG Grants SCHU 2580/1 and SCHU 2580/2). The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

References

Mohamed Abdalla, Krishnapriya Vishnubhotla, and Saif Mohammad. 2023. What makes sentences semantically related? a textual relatedness dataset and empirical study. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 782–796, Dubrovnik, Croatia. Association for Computational Linguistics.

Anna Aksenova, Ekaterina Gavrishina, Elisei Rykov, and Andrey Kutuzov. 2022. RuDSI: Graph-based word sense induction dataset for Russian. In *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing*, pages 77–88, Gyeongju, Republic of Korea. Association for Computational Linguistics.

David Alfter, Rémi Cardon, and Thomas François. 2022. A dictionary-based study of word sense difficulty. In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, pages 17–24.

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.

Susan Windisch Brown. 2008. Choosing sense distinctions for WSD: Psycholinguistic evidence. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 249–252, Stroudsburg, PA, USA.

Pierluigi Cassotti, Lucia Siciliani, Marco de Gemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023a. 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.

Pierluigi Cassotti, Lucia Siciliani, Lucia C. Passaro, Maristella Gatto, and Pierpaolo Basile. 2023b. Wic-ita at EVALITA2023: overview of the EVALITA2023 word-in-context for italian task. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), Parma, Italy, September 7th-8th, 2023*, volume 3473 of *CEUR Workshop Proceedings*. CEUR-WS.org.

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.

Gonzalo Donoso and David Sanchez. 2017. Dialectometric analysis of language variation in twitter. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 16–25, Valencia, Spain.

DWDS. 2021. Digitales Wörterbuch der deutschen Sprache. Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart, hrsg. v. d. Berlin-Brandenburgischen Akademie der Wissenschaften. https://www.dwds.de/. Accessed: 02.02.2021.

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

Adil Fahad, Najlaa Alshatri, Zahir Tari, Abdullah Alamri, Ibrahim Khalil, Albert Y Zomaya, Sebti Foufou, and Abdelaziz Bouras. 2014. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE transactions on emerging topics in computing*, 2(3):267–279.

Mario Giulianelli, Iris Luden, Raquel Fernandez, and Andrey Kutuzov. 2023. Interpretable word sense representations via definition generation: The case of semantic change analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3130–3148, Toronto, Canada. Association for Computational Linguistics.

Serge Kotchourko. 2021. Optimizing human annotation of word usage graphs in a realistic simulation environment. Bachelor thesis, University of Stuttgart.

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.

Sinan Kurtyigit, Maike Park, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2022. DiscoWUG: Discovered Diachronic Word Usage Graphs for German.

Andrey Kutuzov, Mariia Fedorova, Dominik Schlechtweg, and Nikolay Arefyev. 2024. Enriching word usage graphs with cluster definitions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6189–6198, Torino, Italia. ELRA and ICCL.

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.

Jey Han Lau, Paul Cook, Diana McCarthy, Spandana Gella, and Timothy Baldwin. 2014. Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 259–270, Baltimore, Maryland. Association for Computational Linguistics.

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

Olga Majewska, Diana McCarthy, Jasper J. F. van den Bosch, Nikolaus Kriegeskorte, Ivan Vulić, and Anna Korhonen. 2021. Semantic Data Set Construction from Human Clustering and Spatial Arrangement. *Computational Linguistics*, pages 1–48.

Diana McCarthy, Marianna Apidianaki, and Katrin Erk. 2016. Word sense clustering and clusterability. *Computational Linguistics*, 42(2):245–275.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 279–286, Barcelona, Spain.

Diana McCarthy and Roberto Navigli. 2009. The english lexical substitution task. *Language Resources and Evaluation*, 43(2):139–159.

Hermann Paul. 2002. *Deutsches Wörterbuch: Bedeutungsgeschichte und Aufbau unseres Wortschatzes*, 10. edition. Niemeyer, Tübingen.

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.

Martin Pincus. 1970. A monte carlo method for the approximate solution of certain types of constrained optimization problems. *Operations Research*, 18(6):1225–1228.

Rachel E. Ramsey. 2022. Individual differences in word senses. *Cognitive Linguistics*, 33(1):65–93.

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

Dominik Schlechtweg, Enrique Castaneda, Jonas Kuhn, and Sabine Schulte im Walde. 2021a. Modeling sense structure in word usage graphs with the weighted stochastic block model. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 241–251, Online. Association for Computational Linguistics.

Dominik Schlechtweg, Haim Dubossarsky, Simon Hengchen, Barbara McGillivray, and Nina Tahmasebi. 2021b. DWUG EN: Diachronic Word Usage Graphs for English.

Dominik Schlechtweg, Haim Dubossarsky, Simon Hengchen, Barbara McGillivray, and Nina Tahmasebi. 2022a. DWUG EN: Diachronic Word Usage Graphs for English.

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, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2021c. DWUG DE: Diachronic Word Usage Graphs for German.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2022b. DWUG DE: Diachronic Word Usage Graphs for German.

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. 2021d. 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, and Nikolay Arefyev. 2024a. The lscd benchmark: a testbed for diachronic word meaning tasks.

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*.

Nina Tahmasebi, Simon Hengchen, Dominik Schlechtweg, Barbara McGillivray, and Haim Dubossarsky. 2021. DWUG SV: Diachronic Word Usage Graphs for Swedish.

Benjamin Tunc. 2021. Optimierung von Clustering von Wortverwendungsgraphen. Bachelor thesis, University of Stuttgart.

Warren Weaver. 1949/1955. Translation. In William N. Locke and A. Donald Boothe, editors, *Machine Translation of Languages*, pages 15–23. MIT Press, Cambridge, MA. Reprinted from a memorandum written by Weaver in 1949.

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.

A Datasets

Find an overview of the datasets in Table 5. Figure 4 shows the correspondence between the datasets, annotation rounds and versions.

B Change scores

Find robustness checks for change scores over dataset versions in Figure 5. To calculate the change score of a word between two time periods, T1 and T2, we first determine the frequency of each cluster within these periods. This involves counting how many instances from each cluster appear in T1 and T2, respectively. These counts are then used to create probability distributions p and q, which are the cluster distribution over T1 and T2. The difference between these distributions is measured using the JSD. Results align with those observed when assessing cluster variance in Figure 3. Specifically, in the datasets V1, V2, and V3, introducing 40% of noisy edges results in a weak (0.21-0.40) or moderate (0.41-0.60) correlation. However, when analyzing the `resampled` datasets, the correlation remains very strong (0.81-1.00) for German and Swedish, and strong (0.61-0.80) for English.

Additionally, and similar to the cluster validity evaluation in Section 6.1, we assess the correspondence of change scores to the "ground truth" given by the external gold standard and the final round of annotation. Specifically, we evaluate the accuracy of change labels (binary change) and the Spearman correlation (graded change) by comparing the ground truth labels to those derived from various annotation rounds. The results are shown in Figure 6. The plots on the left use the DWUG DE sense as ground truth, while those on the right use the V3 version of each dataset as ground truth. In all cases, we observe upward trends, indicating a convergence with more annotation rounds.

C Semantic change scores convergence

One of the most popular uses of WUGs is the analysis of the semantic change of words. Semantic change scores from WUGs are usually computed using `COMPARE`, the `EARLIER` or `LATER` measures (Schlechtweg et al., 2018), or the Jensen-Shannon Distance (JSD), i.e.,

$$\sqrt{\frac{D(P\,||\,M) + D(Q\,||\,M)}{2}}$$

between the probability distributions of clusters in different historical periods P, Q, where D is the Kullback-Leibler Divergence and $M = \frac{(P+Q)}{2}$ (Lin, 1991; Donoso and Sanchez, 2017; Schlechtweg et al., 2020). Words that have changed meaning will have a higher JSD.

By examining the entropy of the cluster probability distributions, we can categorize words into low-entropy (likely monosemous) and high-entropy (likely polysemous) groups and explore how these groups behave over time in the annotation process. We conduct an experiment on the `resampled` datasets to investigate how different word groups (low-entropy vs. high-entropy) converge towards their final semantic change scores as more data is considered. Our hypothesis is that low-entropy words need less annotation (Kotchourko, 2021). For each WUG, we simulate an incremental annotation process through multiple rounds, iteratively adding a fixed percentage of edges from the annotation to the existing graph in steps. Each step incorporates an additional 5% of edges relative to the previous step, resulting in a sequence of graphs with progressively more edges. Then, we cluster each graph in the sequence and compute the cluster probability distribution and change score. Figure 7 shows the absolute error of change scores obtained on different proportions of data compared to the change score obtained on the full portion of data for two categories, i.e., low- (≤ 0.2) and high- (≥ 0.8) entropy words.[12]

For all words, we observe that semantic change scores can be approximated well in the early stages, with a difference of less than 0.25 dropping below 0.10 after introducing only 20% of the edges. However, there are clear differences between the word groups: words with higher entropy (polysemous words) tend to show a higher error compared to their final change score while words with lower entropy (monosemous words) converge much more quickly. For low-entropy words, we can achieve a perfect approximation of the semantic change score after introducing only 40% of the edges, whereas high-entropy words require more data for accurate approximation.

This suggests that low-entropy words, which

[12] Entropy is calculated considering the cluster obtained using the graph with all the annotated edges. It is important to note that this ground truth entropy (computed on the final graph, i.e., with all the available annotations) is not available in the early stages of the actual annotation process.

Dataset	LGS	|T|	N/V/A	|U|	t_1	t_2	Reference
DWUG	DE	50	34/14/2	178	1800–1899	1946–1990	Schlechtweg et al. (2021d)
DWUG	EN	46	40/6/0	191	1810–1860	1960–2010	Schlechtweg et al. (2021d)
DWUG	SV	44	32/5/7	171	1790–1830	1895–1903	Schlechtweg et al. (2021d)
DiscoWUG	DE	75	39/16/20	49	1800–1899	1946–1990	Kurtyigit et al. (2021)
DWUG Sense	DE	24	16/7/1	50	1800–1899	1946–1990	Schlechtweg (2023)
DWUG `resampled`	DE	15	10/4/1	50	1800–1899	1946–1990	this paper
DWUG `resampled`	EN	15	14/1/0	50	1810–1860	1960–2010	this paper
DWUG `resampled`	SV	15	10/3/2	50	1790–1830	1895–1903	this paper

Table 5: Version-independent dataset statistics. LGS = language, $|T|$ = no. of target words, N/V/A = no. of nouns/verbs/adjectives, $|U|$ = avg. no. uses per word, t_1/t_2 = time periods for first and second corpus.

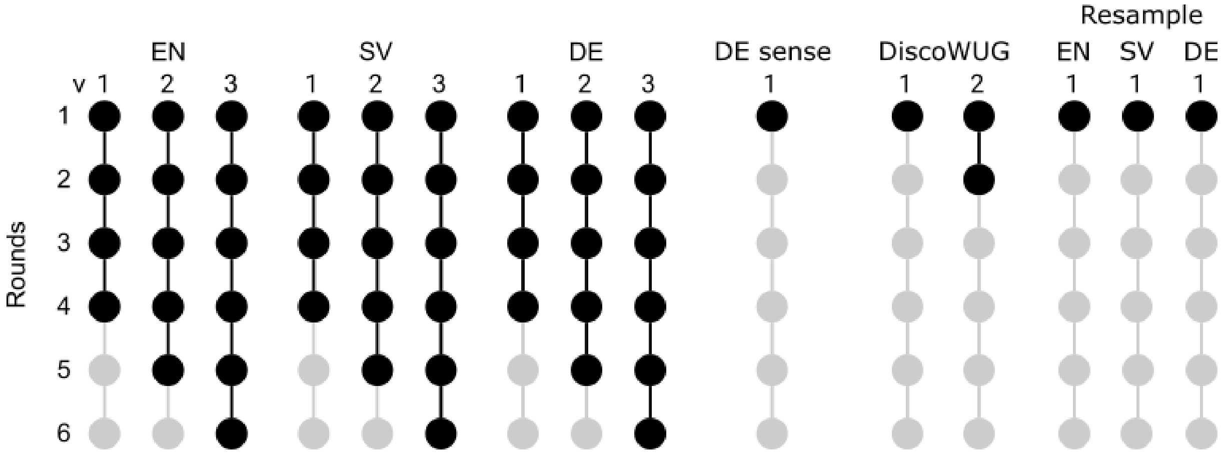

Figure 4: Overview of the datasets, annotation rounds and versions. Black dots represent manual annotation in the given round, while gray dots represent no annotation in that round.

are likely to be monosemous, tend to stabilize in their semantic change scores more rapidly than high-entropy words, which are likely to be polysemous.

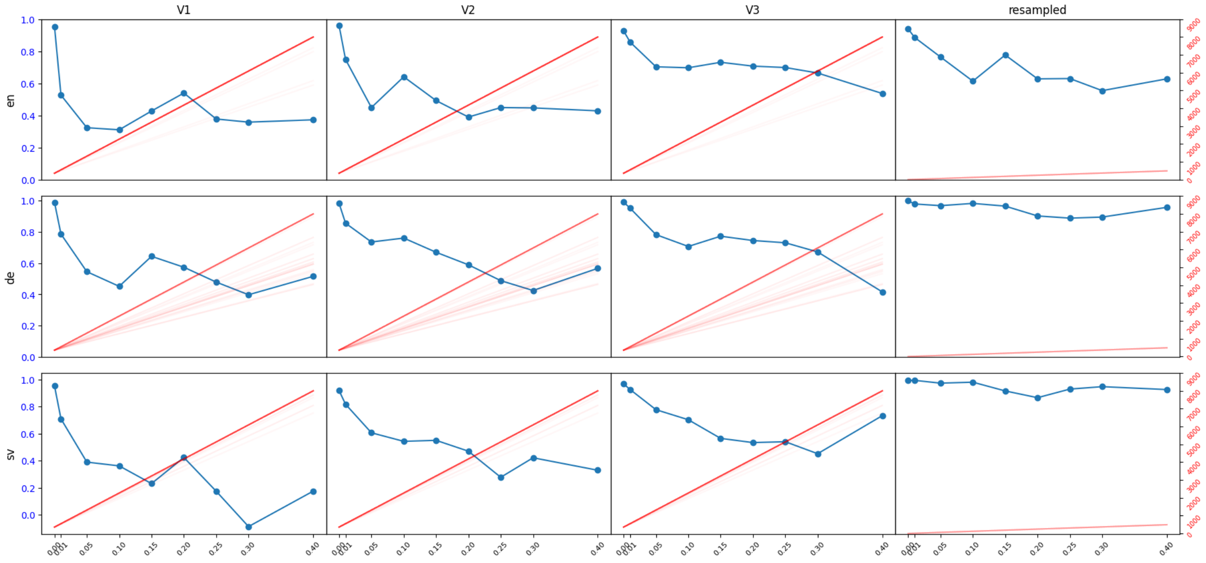

Figure 5: Robustness - Spearman correlation of change scores computed with respect to increasing percentages of noisy edges. The right y-axis (in *red*) shows the raw number of noisy edges. The x-axis shows the percentage of perturbed edges.

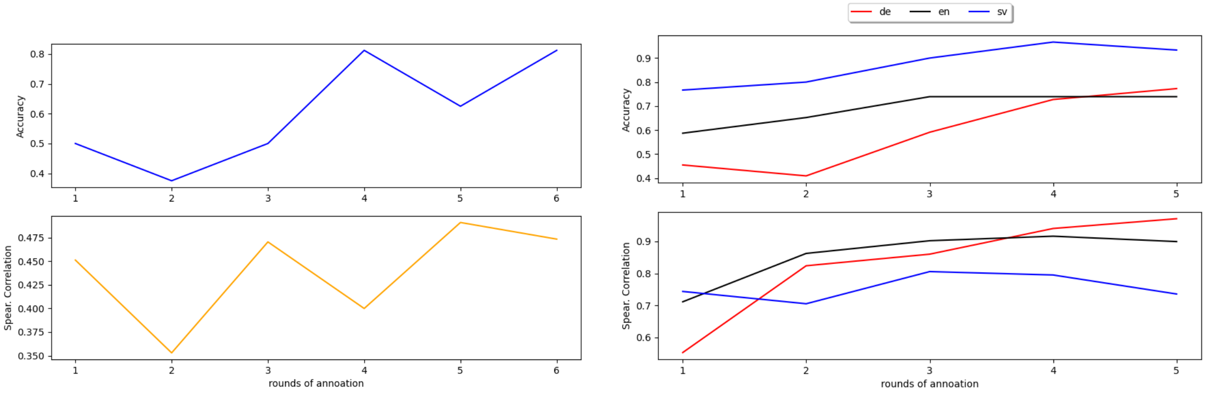

Figure 6: Left: Accuracy (binary change) and Spearman correlation (graded change) of DWUG DE change scores after each round of annotation compared to DWUG DE Sense. Right: Accuracy and Spearman of DWUG DE/EN/SV change scores after each round of annotation compared to V3. Only lemmas that were annotated in each round are included.

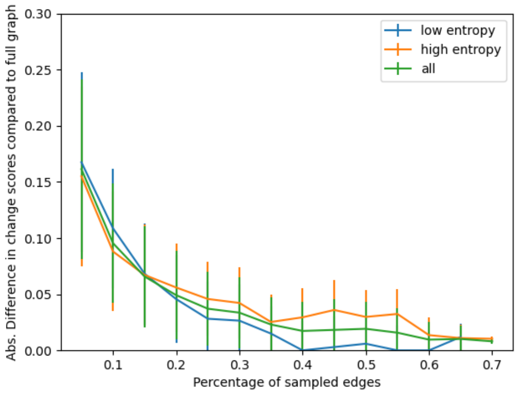

Figure 7: Approximation of final semantic change score in `resampled` datasets considering increasing percentage of edges. The y-axis shows the absolute difference in change score computed at each step.