AmDi - Ambiguous Words Diachronic Dataset

Felix Thielen, Kai Kugler
Trier Center for Digital Humanities, Computerlinguistik
Universität Trier
{thielenf, kuglerk}@uni-trier.de

Abstract

Two fundamental tasks in computational linguistics are Lexical Semantic Change Detection and Word Sense Disambiguation. Both commonly rely on large annotated datasets. Most available datasets cover only one of two areas: diachronic corpora used for Semantic Change Detection, or synchronic datasets for Word Sense Disambiguation. To address this gap, the AmDi dataset is introduced as a German-language resource that supports a more fine-grained diachronic analysis of word meanings, while also enabling the investigation of embeddings generated with corresponding models, as well as providing a foundation for Word Sense Disambiguation tasks.

Keywords: lexical semantic change detection, word sense disambiguation, corpus, dataset

1. Introduction

Lexical Semantic Change (LSC) involves the evolution of word meanings over time, including the acquisition of new senses, loss of existing ones, or subtle shifts in usage. In computational linguistics, early approaches used static embeddings trained on temporally segmented corpora (Hamilton et al., 2016), while recent methods leverage contextualized representations from large language models (LLMs) to capture semantics at the individual word usage level (Tahmasebi et al., 2021). Current research often focuses on Graded Change Detection (GCD), which quantifies the degree of semantic change of target words over time (Schlechtweg et al., 2021). Despite advancements in embedding-based methods, most studies lack resources that link contextualized usage data with explicit sense inventories necessary for sense-level interpretation and evaluation.

Word Sense Disambiguation (WSD) is a canonical task that aims to identify the sense of polysemous[1] words in context drawn from an underlying sense inventory. Supervised approaches have been shown to yield the best results on this task, achieving near perfect accuracy in certain conditions (Ballout et al., 2024). This task therefore also relies on the availability of annotated datasets.

To address both applications, we introduce a German dataset for Lexical Semantic Change Detection (LSCD) and WSD. The dataset includes sense-annotated usages of ambiguous words, sampled by decade from 1901 to 2018, ensuring consistent temporal coverage. Each usage instance is linked to the semantic lexicon GermaNet (Hamp and Feldweg, 1997), providing interpretable sense identifiers that enable direct comparison between automatically induced senses and lexicon-based semantic relations. The dataset is balanced across target words and decades, ensuring comparability and statistical robustness for both sense- and usage-based analyzes.

In addition to the dataset, we present baseline approaches for both LSCD and WSD. We conduct two experiments: One uses fine-tuning of three German-language transformer-based models for classification (see Section 4.1.1), and a few-shot prompting approach (see Section 4.1.3), leveraging GermaNet to generate sense information, which is included in the prompt. We evaluate six models to investigate the extent of their inherent capabilities for this task.

In addition to these supervised approaches, we also implement unsupervised and representation-based methods. For WSD, we apply unsupervised clustering of word usages (see Section 4.1.2). For LSCD, we employ both static and contextualized word representations: we train static word vectors on diachronic corpus segments (see Section 4.2.1) and use contextualized embeddings from pretrained language models and finetuned models for lexical disambiguation (see Section 4.2.2), to analyze and quantify sense shifts over time.

2. Related Work

2.1. Datasets

Most WSD datasets (Senseval-2 (Edmonds and Cotton, 2001), Senseval-3 (Snyder and Palmer, 2004), SemEval-2015 (Moro and Navigli, 2015)) contain many annotated instances but are drawn from single time periods, making them unsuitable for diachronic analysis. Conversely, LSCD corpora like SemEval-2020 Task 1 (Schlechtweg et al., 2020) are time-stratified but lack aligned sense annotations; they focus on detecting usage

[1] We use 'polysemy' to refer to related senses as well as rarer cases of accidental homonymy.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11926–11941
11–16 May 2026.

	AmDi	TüBa-D/Z	GLASS	SemEval-2020	DWUG DE
samples	22,230	18,412	2,038	114,087	1,200
tokens	637,871	470,980	25,929	3,734,441	40,841
lexemes	26	115	153	48	24
timespans	14	5	–	2	2
senses	80	304	210	–	90
task	WSD+LSCD	WSD	WSD	LSCD	LSCD

Table 1: Comparison of **AmDi** with established German WSD and LSCD datasets.

changes rather than mapping to fixed sense inventories like GermaNet. In DWUG DE Sense (Schlechtweg et al., 2021), senses derive from aggregated human annotations without links to external resources. For German, sense-annotated corpora (TüBa-D/Z (Henrich, 2015), GLASS (Miller et al., 2016), DeWSD (Broscheit et al., 2010)) do not encode diachronic variation. TüBa-D/Z is unevenly distributed and its timespans reflect newspaper publication years rather than being designed for LSCD analysis. To contextualize AmDi, we compare it against established datasets for WSD and LSCD (see Table 1).

2.2. Word Sense Disambiguation

Most current WSD systems are based on transformer architectures. Two common methods leverage contextual embeddings as features or fine-tune models on annotated data. The best-performing systems often integrate external knowledge sources, such as knowledge graphs or glosses, to enhance disambiguation accuracy (Bevilacqua and Navigli, 2020; Wang and Wang, 2020; Zhang et al., 2022). However, most performance gains stem from obtaining more sense-annotated data rather than novel methodological approaches (Bevilacqua et al., 2021). Recently, experiments with zero-shot prompting using instruction-tuned models have shown promising results, although they have not yet reached state-of-the-art accuracy (Basile et al., 2025).

2.3. Lexical Semantic Change Detection

LSCD approaches vary in the meaning representation used for lexical items and the method for detecting change. A useful high-level distinction is between *form-based* and *sense-based* representations (Tahmasebi et al., 2021).

Form-based methods represent words as single vectors aggregated across usages at a point in time and detect change by comparing these representations between time periods. Common representations include count-based distributional vectors, static neural embeddings (Mikolov et al., 2013; Pennington et al., 2014) trained separately per period (Hamilton et al., 2016), and dynamic embedding techniques. Form-based LSCD methods are efficient but hard to interpret, hindering sense-level attribution.

Sense-based approaches model individual senses (Giulianelli et al., 2020) or usages (Schlechtweg et al., 2021) instead of merging them into one vector. They enable semantic change analysis using contextualized embeddings from models like BERT (Devlin et al., 2019) or explicit sense annotation or word-sense disambiguation (Tahmasebi et al., 2021). Such methods can capture sense emergence, loss, shift, or gradual change by comparing sense clusters or distributions over time (Tahmasebi and Risse, 2017; Tang et al., 2023; Schlechtweg et al., 2024), but they rely on accurate sense labels.

Evaluation remains a major challenge in LSCD. Most studies use intrinsic metrics without standardized, sense-level gold data. SemEval-2020 Task 1 (Schlechtweg et al., 2020) advanced comparability with graded annotations and shared tasks, while recent work links LSCD with WSD through joint modeling of Word-in-Context and sense induction.

3. The AmDi Dataset

The AmDi dataset (Thielen and Kugler) bridges the existing gaps by providing GermaNet-based sense annotation across multiple historical timespans, thus combining the strengths of the LSCD and WSD paradigms and providing a German resource for both tasks.

3.1. Data Acquisition

We collected textual data in German from the DWDS (Digitales Wörterbuch der Deutschen Sprache) corpus (DWDS) using its public web interface and JSON API. Our target words, including orthographic variants (inflected forms), were queried across three corpora: `public`, `blogs`, and `wikipedia`. The majority of the data was sampled from the `public` corpus, covering consecutive decades from 1901 to 2018. To include more recent and genre-diverse texts, we additionally queried `blogs` (1995–2014) and `wikipedia` (2006–2023). Each query retrieved full sentences containing the target word. The responses were

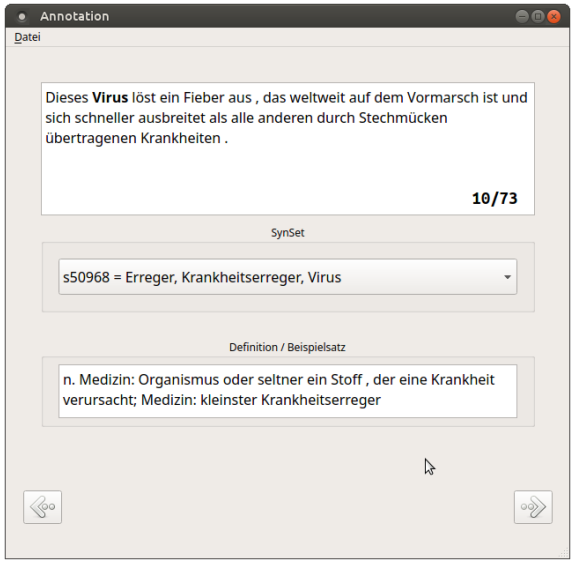

Figure 1: The GUI of the annotation tool displays the target word in context, the list of synsets to choose from, and the paraphrase or example sentence.

parsed to identify token positions, normalize spacing and punctuation, and store each sentence along with metadata such as the target word and a unique identifier in JSON Lines format. This resulted in a structured dataset suitable for subsequent lexical-semantic analysis and embedding-based clustering.

3.2. Annotation

The data was annotated in two steps. In the first step, seventeen mostly high-frequency German polysemous nouns were selected as target words. Based on `wortschatz-leipzig.de` (Goldhahn et al., 2012) the frequency classes of the target words (in corpus `deu_news_2012_3M`) rank between 7 ("Bank" with frequency 15.306) and 14 ("Schimmel" with frequency 89). On the one hand, the selection of these words was motivated by existing data sets: *Schimmel* and *Decke* (and also *Gimpel*, a polyseme to the target word *Tor*) occur in WIC-TSV (Breit et al., 2021), *Fuß* and *Tor* are taken from TüBa-D/Z (Henrich, 2014) and *Schimmel*, *Zeitung*, *Atmosphäre* and *Block* are German translations of the polysemes used in RAW-C (Trott and Bergen, 2021). On the other hand, some target words were inspired by research interest in a specific topic, such as word usage during the coronavirus crisis (*Maske*, *Virus*).

Base target words annotation Sixteen native German-speaking students assigned GermaNet (Hamp and Feldweg, 1997) synsets to base target words. Data was divided into 12 batches (150 samples each), with each sample randomly assigned to two annotators, ensuring even coverage across words and time periods. One subset of 100 samples received four annotations. Annotators chose their workload, resulting in samples annotated 1–4 times (Table 2). A custom GUI tool (see 1) displayed sentences from DWDS together with possible synsets and definitions. Annotators selected the appropriate synset or 'None' if no sense matched. The annotators were computational linguistics students familiar with the GermaNet synsets. In this first step, 10,020 instances were annotated with 79% inter-annotator agreement.

samples	w agr.	w/o agr.	total	agr.
single	19,444	–	19,444	
double	2,631	699	3,330	79%
triple	93	28	121	77%
quadruple	62	23	85	73%
total	22,230	750	22,980	

Table 2: Number of samples (annotated from 2 to 4 times) with (100%) and without agreement.

Annotation of additional samples The dataset was expanded by randomly selecting references to 57 synonymous words across the 14 time periods from DWDS. This expansion targeted less frequent senses of the 17 base target words. For example, *Flügel* (wing) rarely appears as *Parteiflügel* (party wing), so examples for the compound word were added. Some of the additional words are also synonyms: Klappe (flap) has a colloquial meaning for the human mouth. A synonym is *Schnabel* (beak), which also refers to the mouthparts of birds. A native German speaker annotated the data semi-automatically for unambiguous cases (e.g., *Rotorblatt*) and manually for ambiguous words (e.g., *Kasse* with six candidates). Data from both annotation steps were then merged.

Reduction to full agreement We decided to only consider sentences with full agreement for the final AmDi dataset. The intention is to prioritize quality over quantity at the cost of losing roughly 3% of our total data (see Table 2).

3.3. Subsets

For the subsets 'small' and 'tiny', we filtered out small classes, reassigning these samples to the residual class 'None'. For every distribution as a set of class counts $C_{lex} = \{c_1, c_2, \ldots, c_n\}$ where c_i is the number of samples belonging to synset i, we compute μ and σ and define the cutoff threshold $T = \max(1, \mu - k \cdot \sigma)$. We keep the synset i if

$c_i \geq T$ and reassign its samples if $c_i < T$. After reassignment, a lexeme distribution is discarded if fewer than two non-'None' classes remain. We choose thresholds $k = 1.0$ and $k = 0.5$ to get our subsets 'small' and 'tiny' (see Table 3). For a comparison of the distributions (synsets per lexeme), see Figures 5, 6 and 7 in the appendix.

	original	small	tiny
samples	22,230	21,363	17,810
tokens	637,871	608,562	504,080
lexemes	26	24	19
synsets	80	69	46
threshold		1.0	0.5

Table 3: Number of samples in the original dataset and after reassigning small classes and discarding distributions with less than two classes.

4. Experiments

4.1. Word Sense Disambiguation

4.1.1. Fine-Tuning

We trained the models gBERT-large (Chan et al., 2020), a German model based on BERT, XLM-RoBERTa-large (Conneau et al., 2020), a multilingual RoBERTa-model, and XL-LEXEME (Cassotti et al., 2023), a WiC (Word-in-Context) pretrained model, to classify synset labels. To this purpose, we converted our dataset into stratified 'huggingface' datasets with a predefined train/test split (90%/10%). Three different experiments were conducted: one involving training a classifier using only the sentence ('context'), and two others combining the target word with the sentence. In one experiment, we marked the token in the sentence ('marked') and in the other, we added the token embedding as an additional input for training ('pair').

4.1.2. Clustering

We performed unsupervised clustering for each lexeme across multiple embedding representations. We considered embeddings derived from standard transformer-based language models (gBERT-large, XLM-RoBERTa-large), as well as embeddings obtained from LLMs designed for text embedding (e.g., Qwen3-Embedding-8B). In addition, we included the best-performing fine-tuned BERT-type models trained on the AmDi and TüBa-D/Z (TüBa-D/Z) corpora.

For each lexeme, we applied three clustering algorithms - k-means (MacQueen, 1967), agglomerative hierarchical clustering (Kaufman and Rousseeuw, 1990), and HDBSCAN (Campello et al., 2013) — to capture different structural assumptions in the data.

To visually inspect the clustering results, we projected the high-dimensional embeddings into two dimensions using both t-SNE (van der Maaten and Hinton, 2008) and UMAP (McInnes et al., 2018).

Quality of clusters Clusters were then evaluated against the gold standard synset labels provided in the sense-annotated datasets. By mapping predicted clusters to these labels, we quantified the extent to which embeddings encode the semantic distinctions captured in the annotation. Subsequently, we computed a set of metrics—including Adjusted Rand Index (ARI) (Hubert and Arabie, 1985), Normalized Mutual Information (NMI) (Strehl and Ghosh, 2002), Purity, and F1-scores to provide a quantitative assessment of clustering performance and facilitate comparisons across models and methods.

ARI measures clustering agreement with the gold standard, corrected for chance, ranging from -1 (anti-correlation) to 1 (perfect agreement). It is sensitive to both cluster assignments and cluster count, providing a robust global measure. NMI quantifies the information shared between predicted and gold clusterings on a $[0, 1]$ scale. Being less sensitive to cluster count and scale-invariant, NMI enables comparison across lexemes with varying numbers of senses.

Purity measures the fraction of each cluster belonging to its dominant gold label, with global purity as the weighted average in $[0, 1]$:

$$\text{Purity}(U, V) = \frac{1}{n} \sum_{j=1}^{C} \max_i |U_i \cap V_j|$$

While intuitive, purity increases with cluster count (trivial single-item clusters achieve purity = 1).

Since cluster IDs are arbitrary, we use the Hungarian algorithm (Kuhn, 1955) to find optimal one-to-one mapping π between predicted clusters and gold labels, maximizing:

$$\text{Accuracy}(U, V) = \frac{1}{n} \sum_{j=1}^{C} |U_{\pi(j)} \cap V_j|$$

After optimal assignment, we compute Macro-F1 (averaging per-class scores) and Accuracy (Micro-F1, aggregating across classes). This suite provides complementary perspectives: ARI and NMI capture global structure, while Accuracy and F1 offer task-oriented evaluation. Macro-F1 weights all senses equally; Micro-F1 favors frequent senses. Together, high purity with low ARI suggests overfragmentation, while high ARI with lower Macro-F1 indicates dominance of frequent senses.

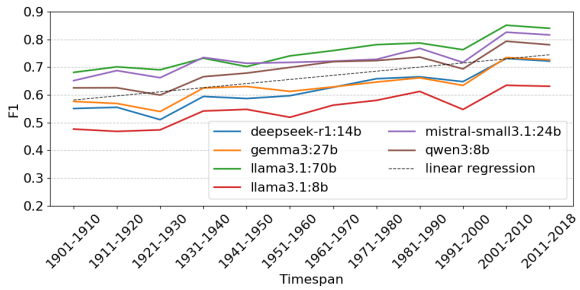

Figure 2: F1 development by decade on the full AmDi dataset. For visual clarity, the data from `blogs`, and `wikipedia` is not plotted as it would cause overlaps with other decades.

Quality of individual instances To analyze clustering quality on the level of individual instances, we computed two per-sample evaluation measures: *sample accuracy* and *sample purity*. Sample accuracy is defined as a binary indicator of whether a sample is assigned to the cluster corresponding to its gold synset. Sample purity measures the degree to which the cluster of a given sample is dominated by its gold synset. For sample i, let $n(\hat{c}_i)$ denote the number of items in cluster $\hat{c}_i$, and $n(y_i, \hat{c}_i)$ the number of items in that cluster whose gold label equals y_i. The sample purity score is defined as:

$$\text{Purity}_i = \frac{n(y_i, \hat{c}_i)}{n(\hat{c}_i)}$$

This yields a graded score in the interval $[0, 1]$: $\text{Purity}_i = 1$ if all items in the cluster share the same gold label as sample i, lower values indicate that the cluster mixes items from multiple gold synsets.

4.1.3. Few-Shot Prompting

We set temperature to zero and instructed the models via system message to respond with the best-matching synset ID, or 'None' if no suitable match can be determined. The prompt was given with a task (i.e., the sentence), a target word, and additional context. The context contained a JSON object generated from GermaNet. For each sense of the target word, it included synset ID, alternate orthographic forms, paraphrases, and up to two levels of hypernyms which were necessary for senses without GermaNet paraphrases. We used the following models: deepseek-r1:14b (DeepSeek-AI et al., 2025), gemma3:27b (Kamath et al., 2025), llama3.1:8b and llama3.1:70b (Grattafiori et al., 2024), mistral-small3.1:24b (Mistral AI Team, 2025), and qwen3:8b (Yang et al., 2025). Both llama3.1 variants and mistral-small3.1 are instruction-tuned. Performance for the full AmDi dataset is shown in Table 4 and Figure 2.

Model	Acc.	Prec.	F1
deepseek-r1:14b	0.655	0.704	0.618
gemma3:27b	0.684	0.762	0.633
llama3.1:70b	**0.773**	0.811	**0.751**
llama3.1:8b	0.562	0.604	0.544
mistral-small3.1:24b	0.757	**0.826**	0.729
qwen3:8b	0.729	0.803	0.695

Table 4: Few-Shot Prompting performance on the full AmDi dataset.

4.2. Semantic Change Detection

4.2.1. Static Word Embeddings

To investigate diachronic semantic change, we trained static word embeddings independently for multiple consecutive timespans. We applied two complementary approaches: per-timespan embeddings with post hoc alignment (Hamilton et al., 2016) and soft temporal regularization.

Per-Timespan Embeddings For each timespan t, we trained a Word2Vec (Mikolov et al., 2013) model producing vectors $\mathbf{v}_w^{(t)}$ for each target word w. Since independently trained embeddings reside in separate vector spaces, we aligned them to a reference timespan using *Orthogonal Procrustes*, finding the optimal rotation R that minimizes:

$$R = \arg\min_{Q \in O(d)} \|Y - XQ\|_F$$

where X and Y are embedding matrices for shared words and $\| \cdot \|_F$ denotes the Frobenius norm. Semantic shift was quantified using cosine distance between consecutive aligned embeddings:

$$\text{shift}_w(t) = 1 - \cos\left(\mathbf{v}_w^{(t)}, \mathbf{v}_w^{(t+1)}\right)$$

Additionally, we computed Jaccard similarity between top-k nearest neighbor sets across timespans. This provides a complementary measure of semantic drift by tracking changes in contextual associations rather than vector position alone.

Soft Temporal Regularization To stabilize embeddings and reduce stochastic drift, we implemented *soft temporal regularization*. For timespan $t > 1$, embeddings were initialized with the previous timespan's vectors $\mathbf{v}_w^{(t-1)}$ and fine-tuned on the current timespan's data. This introduces an implicit regularization term in the objective:

$$L = L_{\text{skipgram}} + \lambda \sum_w \|\mathbf{v}_w^{(t)} - \mathbf{v}_w^{(t-1)}\|^2$$

where L_{skipgram} is the standard skip-gram loss and λ controls the strength of temporal smoothing. This encourages embeddings to remain close to previous representations unless the current context strongly indicates a semantic change.

Quantifying Semantic Shift For both approaches, semantic change for each target word was measured as the cosine distance between embeddings from consecutive timespans. Words missing from a timespan were treated as absent, ensuring robustness to incomplete observations.

Long-range Semantic Shift To complement the pairwise temporal comparisons between adjacent timespans, we performed a long-range semantic shift analysis across all available historical periods. The goal of this analysis was to measure how word meaning diverges as the temporal distance between corpora increases. For each timespan, we used the corresponding Word2Vec model trained on the subset of data from that period. We then computed semantic similarity between all pairs of timespans, not only adjacent ones. For every lexeme w, we calculated the cosine similarity between embeddings of in two models which quantifies directional similarity in the shared vector space and Jaccard similarity between the (k)-nearest neighbor sets of w. This metric captures contextual stability by comparing semantic neighborhoods. For each lexeme, these values were computed for all combinations of timespans, producing a matrix of cosine and Jaccard similarities across temporal distances. We aggregated the results across all lexemes to obtain the average similarity as a function of temporal distance.

4.2.2. Contextualized Embeddings

The contextualized embeddings used for word sense clustering (Section 4.1.2) also served as the foundation for detecting lexical semantic change. Two complementary approaches were applied to quantify temporal shifts in word meaning:

Cosine-based change detection The method of Martinc et al. (2020) was applied to the dataset by computing time-specific mean embeddings for each lexeme. Since the dataset contains samples across consecutive timespans, contextualized embeddings could be used directly to calculate these means. Semantic change between consecutive timespans t_1 and t_2 was quantified using cosine distances between the averaged embeddings, providing a straightforward measure of lexical shift over time:

$$\text{shift}_w(t_1 \to t_2) = 1 - \frac{\mu_{w,t_1} \cdot \mu_{w,t_2}}{\|\mu_{w,t_1}\| \, \|\mu_{w,t_2}\|}$$

Continously Evolving Embeddings We also employed a pipeline of *continuously evolving embeddings* (Horn, 2021) using the contextualized word representations from the pretrained models. For each lexeme w and timespan t, we computed an *exponentially weighted running mean* (Finch, 2009) of its embeddings $\mathbf{e}_{w,t}^{(i)}$:

$$\mu_{w,t}^{(i)} = \alpha \, \mathbf{e}_{w,t}^{(i)} + (1 - \alpha) \, \mu_{w,t}^{(i-1)}$$

where $\alpha \in (0, 1]$ is a smoothing factor. This allows efficient incremental averaging of embeddings without storing all individual occurrences. Semantic shift between consecutive timespans was quantified using cosine distance between the running mean embeddings.

Model	Dataset	mark	pair	sent.
gbert-large	AmDi	**0.781**	0.774	0.737
gbert-large	AmDi s.	0.824	0.833	0.797
gbert-large	AmDi t.	0.853	0.857	0.841
gbert-large	TüBa-D/Z	**0.906**	0.888	0.583
xl-lexeme	AmDi	0.749	0.750	0.669
xl-lexeme	TüBa-D/Z	0.793	0.693	0.402
xlmr-large	AmDi	0.723	0.703	0.701
xlmr-large	TüBa-D/Z	0.773	0.772	0.411

Table 5: Results (F1 avg weighted) of the fine-tuning Experiments on different inputs and subsets.

4.3. Reproducibility and Code Availability

All experimental procedures and model training protocols are implemented using open-source tools including scikit-learn (Buitinck et al., 2013), Gensim (Rehurek and Sojka, 2011), the HuggingFace Transformers library (Wolf et al., 2020) and Ollama (Ollama Team, 2025). The AmDi dataset is available in all versions, including metadata and dataset statistics, at https://gitlab.rlp.net/cl-trier/amdi-dataset. Repositories with code implementations, experimental configurations, and embeddings will also be linked under this address. The fine-tuned models and datasets with test/train-splits will be made available on huggingface.com (see https://huggingface.co/kugler).

5. Results

5.1. Word Sense Disambiguation

5.1.1. Fine-Tuning

The German model (gBERT-large) achieved the best classification results on the marked input on the AmDi dataset. In addition, we trained the models on the subsets 'small' and 'tiny' and on the TüBa-D/Z dataset for comparison. Results are shown in Table 5.

Model	Embedding	ARI	NMI	Purity	Acc	Macro-F1
gbert	raw	0.089	0.124	0.749	0.606	0.399
gbert finet.	raw	0.383	0.393	0.858	0.694	0.586
xlmr	sentence	0.028	0.054	0.704	0.507	0.354
xlmr finet.	sentence	0.432	0.424	0.870	0.716	0.605
qwen3	sense	0.002	0.013	0.676	0.513	0.336
qwen3	sentence	0.189	0.243	0.805	0.594	0.483

Table 6: Overall mean clustering results for different models (gBERT-large, XLM-Roberta-large, Qwen3-Embedding-8B and the fine-tuned BERT-type models), embedding strategies and the best performing clustering method (agglomerative) for the AmDi dataset.

5.1.2. Clustering

Table 6 shows that fine-tuned BERT-type models substantially outperform base models on AmDi. Fine-tuned gBERT-large achieves ARI of 0.383 versus 0.089 for baseline, demonstrating that fine-tuning with marked inputs improves sense separability. Text embedding models perform better with sentence embeddings (Qwen3 ARI: 0.189) than sense-prompted embeddings (0.002), suggesting prompting with synset information does not enhance sense structure. Agglomerative clustering consistently yields best performance. Fine-tuned models show markedly higher scores on AmDi than TüBa-D/Z, while base models and text embedding models remain substantially lower. Embeddings sometimes merge fine-grained GermaNet distinctions (e.g., *Atmosphäre* combines two mood-related synsets; Figure 3). Overall, task-specific fine-tuning significantly improves sense clustering, whereas generic embeddings may not capture fine-grained lexical distinctions.

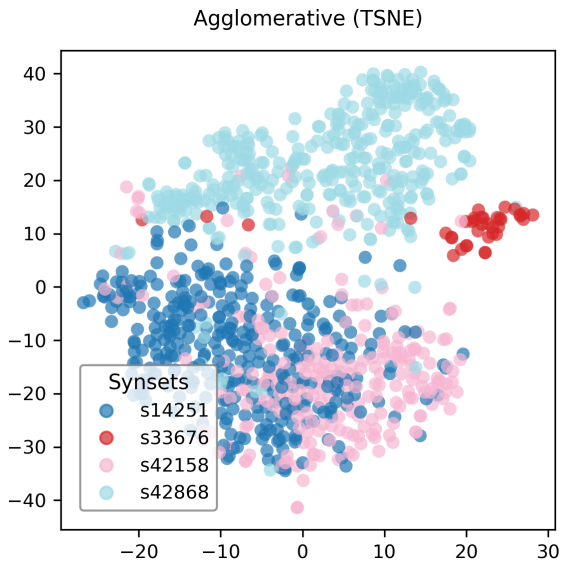

Figure 3: Clustering (agglomerative) of the Qwen3 embeddings for the AmDi lexeme "atmosphäre" (t-SNE).

5.1.3. Few-Shot Prompting

On the full dataset, we achieve respectable overall results, with the best-performing model obtaining an F1 score of 0.751. This suggests that modern LLMs are capable of performing well on WSD tasks without task-specific tuning, given a suitable prompting technique. Although our results indicate a correlation between model size and performance, there are notable exceptions: qwen3:8b, advertised for its strong thinking abilities, significantly overperforms relative to its size. Moreover, mistral-small3.1:24b is only slightly trailing behind llama3.1:70b despite its smaller size. Additionally, we observe a clear recency bias (see Figure 2), which may reflect the distribution of the models' training data. On data sampled from 2001 to 2018, the best two models are easily breaking the 0.8 F1 barrier.

When considering only sentences with two or more annotations, the overall performance of all models improves, reaching a maximum F1 score of 0.839. However, the performance across decades exhibits stronger variance between models and between timespans. We also notice a slight overall decline in performance over time. These effects may be caused by the construction of this subset: since all sentences in AmDi have full annotator agreement, instances with multiple annotations may be more straightforward to disambiguate, as they likely represent less ambiguous or more prototypical word senses. Additionally, the reduced sample size compared to the full dataset makes variance between models and time spans more pronounced.

5.2. Semantic Change Detection

5.2.1. Static Word Embeddings

Word2Vec models trained on successive decades were aligned and compared across time. Cosine and Jaccard similarities showed small differences between neighboring decades, with gradual decay for distant decades (Figure 4), indicating slow, continuous shift rather than abrupt change. Steeper cosine decay compared to Jaccard suggests that

while lexical neighborhoods and contextual associations (reflected in shared neighbors) remain relatively stable, the core vector positions shift more noticeably over time. This aligns with our frequent, stable but ambiguous target lexemes rather than words undergoing major shifts (for per-lexeme cosine and Jaccard similarity see Figures 8 and 8 in the appendix). Nevertheless, results demonstrate that diachronic embeddings capture fine-grained evolution even in homogeneous lexical material.

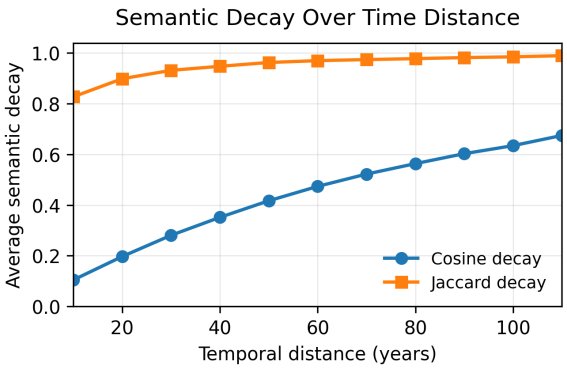

Figure 4: Semantic decay over time for all AmDi target words, Cosine and Jaccard similariry

5.2.2. Contextualized Embeddings

To assess diachronic meaning variation, we computed both cosine-based and running-mean shifts of mean embedding vectors across timespans for all models and embedding strategies.

Cosine-based change detection Qwen3-Embedding-8B showed highest mean change (0.075), though lexical items with largest shifts (e.g., *Gesichtsmaske*) occur in few samples, likely reflecting data sparsity rather than meaning change. Overall differences between decades are small, indicating semantic stability. Sense-level embeddings showed slightly higher sensitivity (0.057) than raw (0.027), CLS (0.017), and sentence (0.013) embeddings. While sense-aware representations capture subtle shifts, observed variation primarily reflects contextual usage differences rather than genuine lexical semantic change.

Continuously evolving embeddings The running-mean analysis produced similar results, with moderate overall change despite some variation (e.g., 0.382 mean change for the Qwen3-Embedding-8B model). Largest shifts for words like *Universität* (university) and *Kurs* (course/direction) appear to reflect contextual rather than semantic change. For DWUG-DE Sense, larger text embedding models identified shifts in *abdecken* (to cover) (0.232) and *Ohrwurm* (earwig) (0.275), aligning with usage-based patterns from Schlechtweg (2023). Averaged across methods, sense embeddings showed strongest change (0.382), followed by raw (0.125), CLS (0.074), and sentence (0.057) embeddings.

Both analyses indicate limited lexical semantic change in the examined corpora. Larger LLM-based models, particularly Qwen variants, exhibit greater sensitivity to contextual variation, but detected differences appear to stem from changing discourse contexts rather than meaning shifts. Finer-grained analysis of contextualized usages or sense distributions is needed for more precise characterization of diachronic lexical dynamics.

6. Future Work

Future work will focus on extending the sense-annotated German dataset along several dimensions.

On the data side, the current version of the dataset contains only nouns. Future extensions will include verbs and adjectives, thereby broadening the lexical and semantic coverage. The target words selection will be inspired by established datasets, ensuring comparability and alignment. Furthermore, coverage could extend to earlier historical periods to enable longer-span diachronic analysis. Cross-linguistic extensions, for instance to other Germanic or European languages, would also make it possible to study semantic change in a comparative perspective.

An extension that is already planned is the inclusion of samples from the DWDS (DWDS) `corona` corpus, which will enable fine-grained analyses of recent lexical usage changes for target words such as *Maske* (mask). In addition, the current data can be repurposed to derive a WiC (Word-in-Context) dataset for German, supporting further research on contextualized meaning representations.

Unresolved samples with annotator disagreement can be exploited to construct usage graphs (Schlechtweg et al., 2018), in which degrees of annotation disagreement are translated into measures of semantic proximity between usages. The existing clustering results, which group semantically related synsets, could potentially inform or refine such graph-based representations.

Finally, since manual annotation time is limited, it may be worthwhile to explore methods for synthetic data expansion, such as semi-automatic generation or model-assisted annotation strategies, to accelerate future dataset growth.

7. Conclusion

We have presented a novel diachronic German dataset comprising synset-annotated with word senses drawn from GermaNet. By focusing on ambiguous target nouns and providing fine-grained sense annotations, this resource enables systematic investigation of lexical semantic change at the sense level.

Our experiments demonstrate that WSD on this dataset is feasible, though the granularity of GermaNet synsets sometimes exceeds what contemporary language models capture in their representations. Fine-tuned BERT-type models achieve substantial improvements in clustering performance. Few-shot prompting yielded similar results, supporting the assumption that modern LLMs have diverse inherent capabilities even without task-specific fine-tuning. At the same time, generic text embedding models show limited ability to distinguish fine-grained senses even when prompted with sense information.

Lexical Semantic Change Detection experiments reveal limited short-term semantic shift for our target words, consistent with their selection as frequent, semantically stable but ambiguous lexemes. However, long-range analyzes using both static Word2Vec embeddings and LLM representations indicate gradual drift in usage patterns over time.

These findings highlight the value of sense-annotated diachronic data to understand how language models encode semantic information. Our dataset provides a foundation for further investigation into how large-scale models capture subtle distinctions in word meaning through contextualization, and how these representations relate to structured lexical resources. Future work could explore whether explicit sense modeling improves the detection of fine-grained semantic evolution, and how different embedding strategies balance stability of core meaning against sensitivity to contextual variation.

8. Ethical Considerations

Data Annotation All annotation work was conducted with transparency regarding the purpose and scope of the project. The annotators were informed in advance about the intended use of the data and the overall research goals. To protect their privacy, personal identifiers were replaced by anonymized annotator IDs. These IDs were subsequently shuffled across work packages, ensuring that individual annotation contributions could no longer be traced back to specific annotators. After completion, the work packages were merged into a single dataset. Annotators were aware of the sources of the sentences to be annotated. The data originate primarily from newspapers and scientific publications, and it was assumed that these materials would not contain harmful or offensive content. No such material was encountered during the annotation process.

Broader Impact The dataset is designed to support research on fine-grained meaning variation and lexical semantic change. However, we acknowledge that methods capable of identifying subtle linguistic differences could, in principle, also be applied to authorship attribution, attitude analysis, or forms of social monitoring. Such potential misuse is a general concern in many areas of text analysis, including sentiment analysis and forensic linguistics. We consider the risk associated with this dataset to be low and assume that the publication and use of the dataset do not pose harm to individuals or groups.

9. Limitations

The current dataset is limited to nouns and texts from the 20th century, restricting coverage of other parts of speech and historical periods. Most data originate from newspapers, scientific texts, fiction, and blogs, which provide a more diverse register but may still underrepresent some informal or spoken language. Although annotation quality was monitored, some inter-annotator disagreement and semantic ambiguity remain. Planned additions may improve topical and temporal coverage.

Methodologically, the unsupervised clustering approach for WSD (see Section 4.1.2) approximates human sense distinctions but may miss subtle contextual nuances. Static word embeddings and contextualized embeddings, while effective for LSCD, are sensitive to domain, genre, and fine-tuning choices, and may reflect biases from their pretrained models.

Finally, there are ethical and practical constraints. The dataset is intended solely for linguistic research and should not be used for authorship attribution or social profiling. Annotator anonymity (see 8) prevents tracking individual contributions, limiting some analyses of reliability. These limitations highlight current boundaries and provide guidance for future extensions and refinements.

10. Bibliographical References

Mohamad Ballout, Anne Dedert, Nohayr Muhammad Abdelmoneim, Ulf Krumnack, Gunther Heidemann, and Kai-Uwe Kühnberger. 2024. FOOL ME IF YOU CAN! an adversarial dataset to investigate the robustness of LMs in word sense disambiguation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5042–5059, Miami, Florida, USA. Association for Computational Linguistics.

Pierpaolo Basile, Lucia Siciliani, Elio Musacchio, and Giovanni Semeraro. 2025. Exploring the word sense disambiguation capabilities of large language models.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864, Online. Association for Computational Linguistics.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Anna Breit, Artem Revenko, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. WiC-TSV: An evaluation benchmark for target sense verification of words in context. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1635–1645, Online. Association for Computational Linguistics.

Samuel Broscheit, Anette Frank, Dominic Jehle, Simone Paolo Ponzetto, Danny Rehl, Anja Summa, Klaus Suttner, and Saskia Vola. 2010. Rapid bootstrapping of word sense disambiguation resources for german. In *Semantic Approaches in Natural Language Processing : Proceedings of the Conference on Natural Language Processing 2010*, Saarbrücken. Universaar.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Ricardo JG Campello, Daniel Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172. Springer.

Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6788–6796. International Committee on Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, and Qihao Zhu et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Philip Edmonds and Scott Cotton. 2001. Senseval-2: Second international workshop on evaluating word sense disambiguation systems. In *Proceedings of Senseval-2*.

Tony Finch. 2009. Incremental calculation of weighted mean and variance.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, and Akhil Mathur et al. 2024. The llama 3 herd of models.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a lexical-semantic net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

Erhard Henrich, Verena; Hinrichs. 2014. Consistency of manual sense annotation and integration into tüba-d/z treebank. *Proceedings of the ThirteenthInternational Workshop on Treebanksand Linguistic Theories (TLT13) : December 12-13, 2014, Tübingen, Germany / Editors: Verena Henrich, Erhard Hinrichs, Daniël de Kok, Petya Osenova, Adam Przepiórkowski*.

Verena Henrich. 2015. *Word Sense Disambiguation with GermaNet*. Ph.D. thesis, University of Tübingen, Germany.

Franziska Horn. 2021. Exploring word usage change with continuously evolving embeddings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 290–297, Online. Association for Computational Linguistics.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*, 2(1):193–218.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, and Alexandre Ramé et al. 2025. Gemma 3 technical report.

L. Kaufman and P. J. Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley.

Harold W. Kuhn. 1955. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97.

James MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297.

Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2020. Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France. European Language Resources Association.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Tristan Miller, Mohamed Khemakhem, Richard Eckart de Castilho, and Iryna Gurevych. 2016. Sense-annotating a lexical substitution data set with ubyline. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 828–835, Portorož, Slovenia. European Language Resources Association (ELRA).

Mistral AI Team. 2025. Mistral Small 3.1: A New Leader in the Small Models Category with Image Understanding Capabilities. https://mistral.ai/news/mistral-small-3-1. Blog post released March 17, 2025.

Andrea Moro and Roberto Navigli. 2015. SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297, Denver,

Colorado. Association for Computational Linguistics.

Ollama Team. 2025. Ollama. GitHub Repository. Version 0.11.3.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Dominik Schlechtweg. 2023. Human and computational measurement of lexical semantic change. *PhD thesis. University of Stuttgart.*

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 Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

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, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021. 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. 2024. The lscd benchmark: a testbed for diachronic word meaning tasks.

Benjamin Snyder and Martha Palmer. 2004. Senseval-3: Third international workshop on the evaluation of systems for the semantic analysis of text. In *Proceedings of Senseval-3*.

Alexander Strehl and Joydeep Ghosh. 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617.

Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen, editors. 2021. *Computational approaches to semantic change*. Number 6 in Language Variation. Language Science Press, Berlin.

Nina Tahmasebi and Thomas Risse. 2017. Finding individual word sense changes and their delay in appearance. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 741–749, Varna, Bulgaria. INCOMA Ltd.

Xiaohang Tang, Yi Zhou, Taichi Aida, Procheta Sen, and Danushka Bollegala. 2023. Can word sense distribution detect semantic changes of words? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3575–3590, Singapore. Association for Computational Linguistics.

Sean Trott and Benjamin Bergen. 2021. Raw-c: Relatedness of ambiguous words–in context (a new lexical resource for english).

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Ming Wang and Yinglin Wang. 2020. A synset relation-enhanced framework with a try-again mechanism for word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6229–6240, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, and Chang Gao et al. 2025. Qwen3 technical report.

Guobiao Zhang, Wenpeng Lu, Xueping Peng, Shoujin Wang, Baoshuo Kan, and Rui Yu. 2022.

Word sense disambiguation with knowledge-enhanced and local self-attention-based extractive sense comprehension. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4061–4070, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

11. Language Resource References

DWDS. *DWDS – Digitales Wörterbuch der deutschen Sprache. Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart*. Berlin-Brandenburgische Akademie der Wissenschaften. Accessed: 2025-10-23.

Hamp, Birgit and Feldweg, Helmut. 1997. *GermaNet*. Version 19.0, Universität Tübingen.

Felix Thielen and Kai Kugler. *AmDi-Dataset*. Computerlinguistik, Universität Trier. Version 'epsilon'.

TüBa-D/Z. *TüBa-D/Z – Tübinger Baumbank des Deutschen / Zeitungskorpus*. Seminar für Sprachwissenschaft, University of Tübingen, 11.0.

A. Dataset Statistics

Timespan	Samples
1901–1910	1978
1911–1920	1806
1921–1930	1779
1931–1940	2116
1941–1950	1706
1951–1960	1321
1961–1970	1313
1971–1980	1423
1981–1990	1452
1991–2000	2122
1995–2014	689
2001–2010	2096
2006–2023	674
2011–2018	1755

Table 7: AmDi (full dataset): Annotated target words per timespan

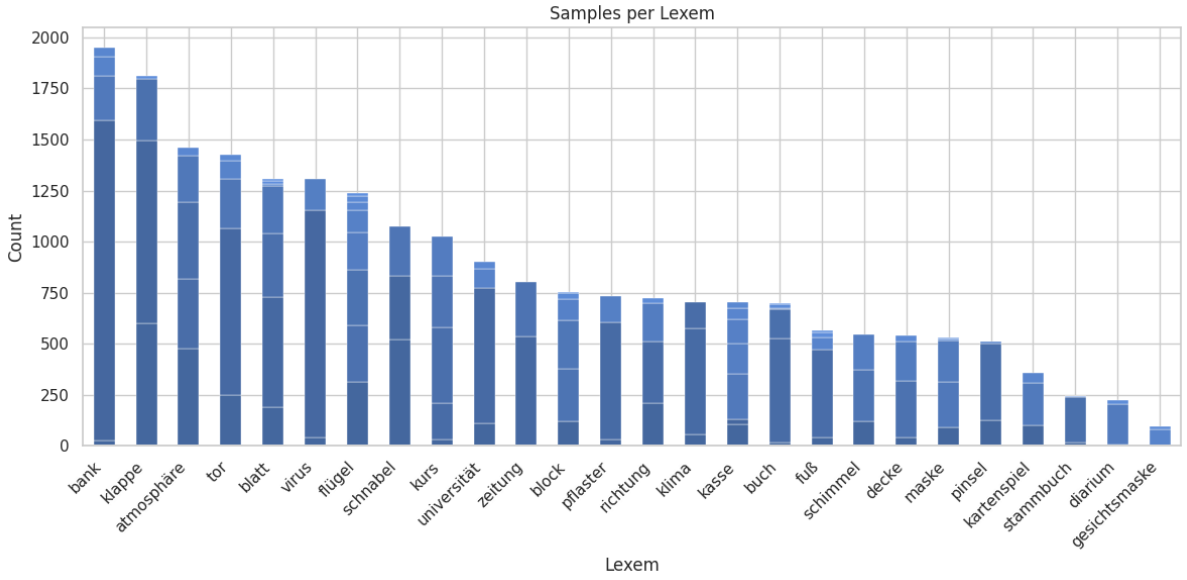

Figure 5: Samples per lexeme (full dataset)

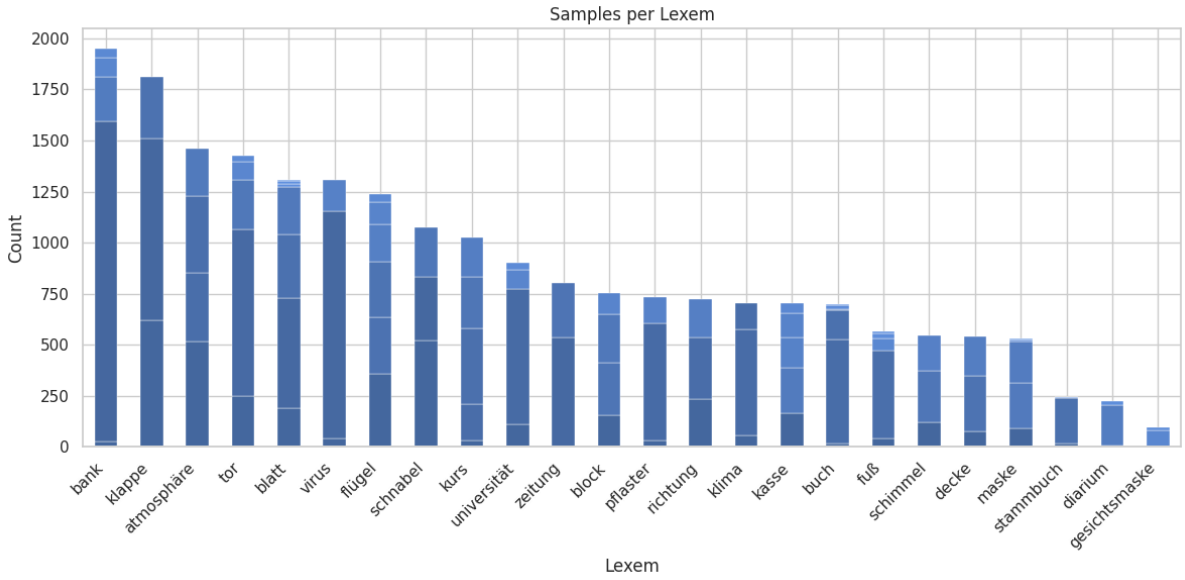

Figure 6: Samples per lexeme (subset small)

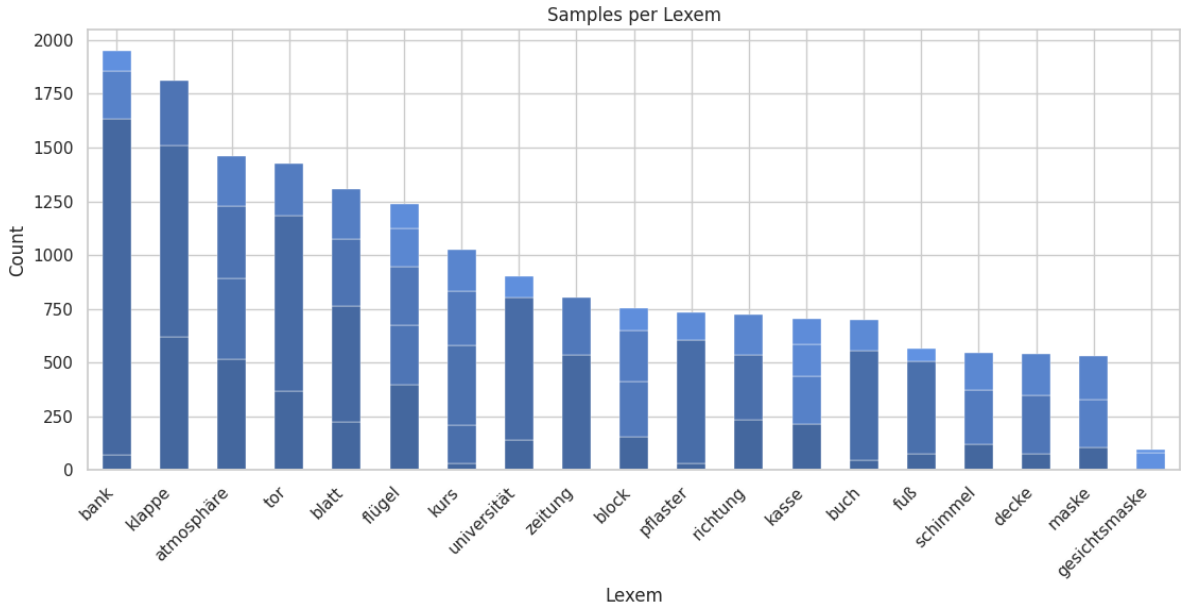

Figure 7: Samples per lexeme (subset tiny)

B. Semantic Drift

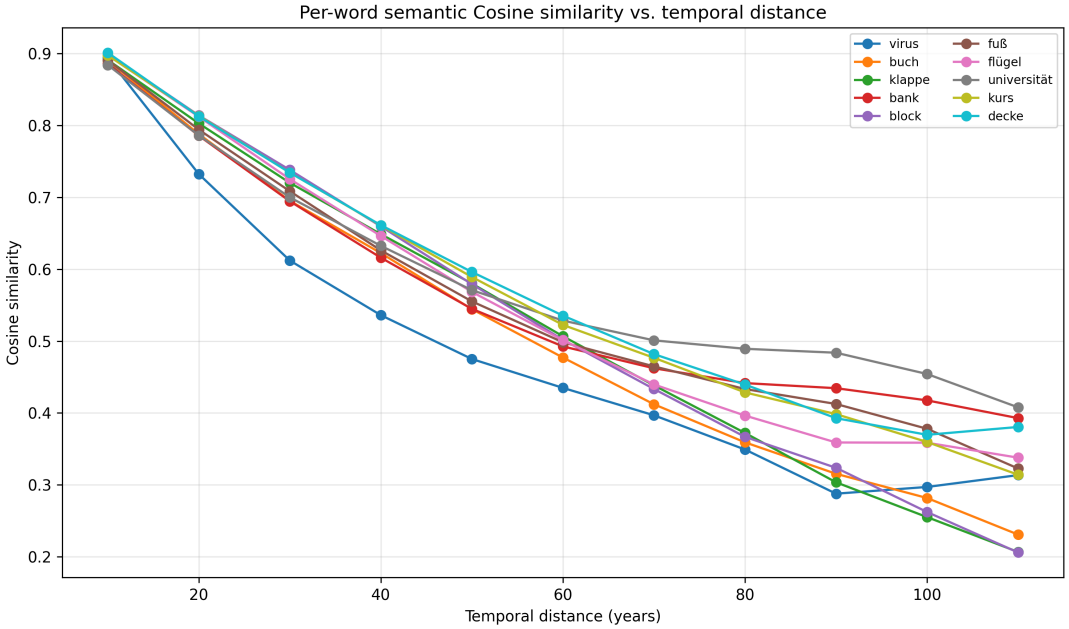

Figure 8: Per lexeme cosine similarity (long-range, temporal regularization)

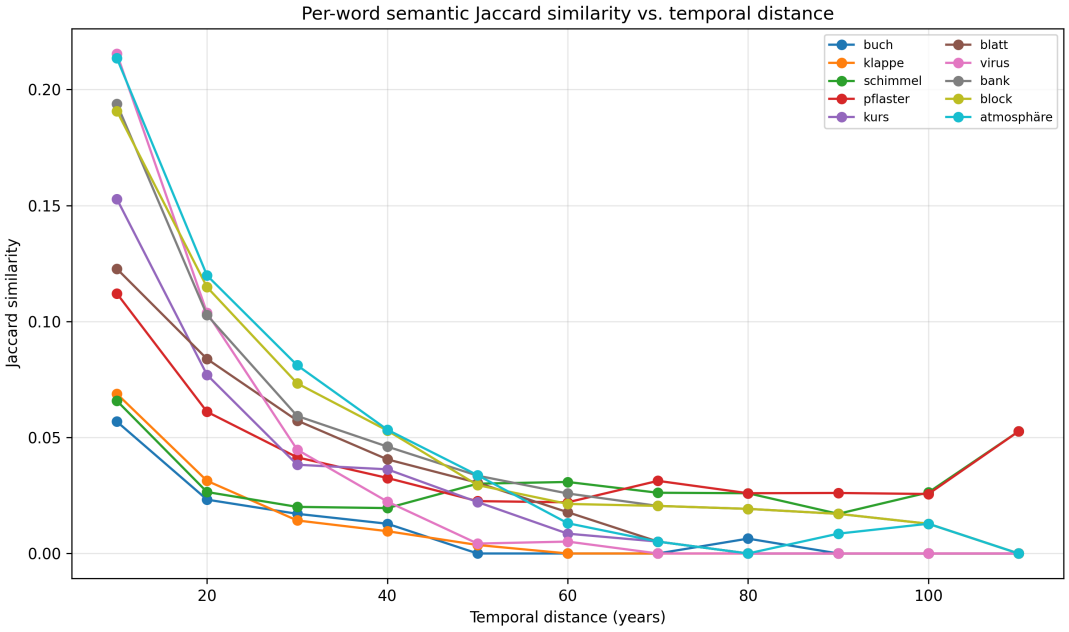

Figure 9: Per lexeme Jaccard similarity (long-range, temporal regularization)