How Do Lexical Senses Correspond Between Spoken German and German Sign Language?

Melis Çelikkol[1] **and Wei Zhao**[2]

Institute for Computational Linguistics, University of Heidelberg[1]
Department of Computing Science, University of Aberdeen[2]
melis.celikkol@stud.uni-heidelberg.de
wei.zhao@abdn.ac.uk

Abstract

Sign language lexicographers construct bilingual dictionaries by establishing word-to-sign mappings, where polysemous and homonymous words corresponding to different signs across contexts are often underrepresented. A usage-based approach examining how word senses map to signs can identify such novel mappings absent from current dictionaries, enriching lexicographic resources. We address this by analyzing German and German Sign Language (Deutsche Gebärdensprache, DGS), manually annotating 1,404 word use–to–sign ID mappings derived from 32 words from the German Word Usage Graph (D-WUG) and 49 signs from the Digital Dictionary of German Sign Language (DW-DGS). We identify three correspondence types: Type 1 (one-to-many), Type 2 (many-to-one), and Type 3 (one-to-one), plus No Match cases. We evaluate computational methods: Exact Match (EM) and Semantic Similarity (SS) using SBERT embeddings. SS substantially outperforms EM overall 88.52% vs. 71.31%), with dramatic gains for Type 1 (+52.1 pp). Our work establishes the first annotated dataset for cross-modal sense correspondence and reveals which correspondence patterns are computationally identifiable. Our code and dataset are made publicly available[1].

1 Introduction

Sign language lexicographers construct bilingual dictionaries by establishing word-to-sign mappings, typically documenting one canonical mapping per word. However, polysemous and homonymous words often correspond to multiple distinct signs across different contexts, yet existing dictionaries may not capture this full range of correspondences. A usage-based approach that examines how word senses map to signs across word usages can identify such novel mappings absent from current dictionaries, enriching lexicographic resources and revealing systematic patterns in how ambiguities transfer across the two modalities.

Lexical ambiguity arises when the meaning of a word changes across different contexts, making its actual sense uncertain until the context is specified. This uncertainty of word sense exists in all languages. Even when two languages use the "same" word, their senses do not align one-to-one. For instance, English *bank* refers to a financial institution or the side of a river, while German *Bank* does not cover the sense of river side. This shows that languages often differ in how senses are encoded within a word and its translation. Identifying such sense correspondence between word translations is crucial in lexicography, language learning and computational linguistics (Hurford et al., 2007; Simatupang, 2007), as this will help lexicographers to build dictionaries.

Identifying sense correspondence becomes more challenging when we compare between spoken and sign languages. Sign languages, as fully developed natural languages operating in the visual-gestural modality, also exhibit lexical ambiguity, where the senses of a sign may align with, deviate from, or partially overlap with the senses of its word translation in a spoken language. This leads to unparallel senses between a word and its sign translation(s) (Johnston and Schembri, 2007; Quer and Steinbach, 2015). Characterizing these correspondence patterns empirically, identifying which patterns exist, and whether computational systems can reliably detect them, remains an open question. For instance, the German word "erlauben" (allow/permit) has three sign translations in DGS, where multiple senses are encoded within a single word form, whereas DGS distributes these senses across three signs. This shows that senses correspond differently across the spoken and sign modalities.

[1] https://github.com/C-Melis/Ambiguity_Resolution_Across_German_Words_and_Their_Sign_Correspondents

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 4: Student Research Workshop, pages 735–746
March 26, 2026 ©2026 Association for Computational Linguistics

While sense correspondence (one-to-many, many-to-one, and one-to-one) across spoken languages has been investigated (Xu et al., 2024; Rahit et al., 2018), little attention has been paid to sense correspondence between spoken and sign languages, as available resources lack semantic annotations required to compare sense correspondence. Although previous studies showed that semantic differences are present across the two modalities (Schulder et al., 2024), existing methods cannot identify how these differences exhibit, especially whether polysemy and homonymy in spoken languages mirror, diverge from, or partially overlap with those in sign languages.

In this project, our aim is to identify the types of sense correspondence between spoken German and German Sign Language (Deutsche Gebärdensprache DGS). To do so, we manually annotate words and their sign correspondence based on two linguistic resources: (i) the German Word Usage Graph (D-WUG) (Schlechtweg et al., 2024), providing word uses, and (ii) the Digital Dictionary of German Sign Language (DW-DGS) (Langer et al., 2024), containing signs through video recordings with unique identifier labels, German translations, and "Erklärung" (the explanation of the sense, the so-called dictionary definition). All words selected from D-WUGs exhibit multiple senses, making them inherently ambiguous and diversifying the types of sense correspondence across the two modalities. By matching German words from D-WUG to their sign translations in DW-DGS, we create a manually-annotated dataset containing three types of cross-modal sense correspondence. Our work makes three key contributions to computational sign language research:

- We provide 1,404 human-annotated mappings from word uses to sign IDs (973 train+val+test_overlap uses, 431 test_no_overlap uses) derived from 32 German words, establishing the first resource for analyzing cross-modal ambiguity correspondence grounded in word usages.

- We identify and characterize three distinct patterns of how ambiguities transfer across modalities: Type 1 (one-to-many, 28.6% of words), Type 2 (many-to-one, 28.6%), and Type 3 (one-to-one, 33.3%), demonstrating that no single pattern dominates cross-modal semantic organization.

- Our semantic similarity method achieves 88.52% accuracy overall, with dramatic improvements for Type 1 (+52.1 pp over exact matching), revealing which types of correspondence are easy to identify and which are not.

2 Related Work

Lexical Ambiguities result from the fact that a single word form can have multiple meanings, primarily through polysemy and homonymy (Klepousniotou, 2002; Haber and Poesio, 2024). Polysemy associates one word with multiple conceptually or historically related senses sharing a common etymological core, while homonymy involves words sharing identical form but having unrelated senses (Fromkin et al., 2010). Similarly, sign languages exhibit lexical ambiguities, with many signs being ambiguous or multifunctional (Quer and Steinbach, 2015; Pfau and Steinbach, 2016). This is due to its distinction from spoken languages: modality-specific properties of sign languages, such as the three-dimensional signing space for establishing referential loci, the simultaneous use of manual and non-manual articulators, role shift enabling perspective adoption, and spatial modification of classifier hand shapes, actively affect their linguistic structure (Johnston and Schembri, 2007). Previous work shows that polysemy and homonymy are present across various sign languages (Gwammaja, 2025; Neubauer, 2024), demonstrating that sign languages are living languages harboring these phenomena actively (Bahan and Dannis, 1996).

To disambiguate the senses of a word or a sign, different disambiguation approaches are applied: For spoken languages, previous work relies on context based on surrounding words and discourse information to resolve lexical ambiguities (Fromkin et al., 2010). For sign languages, previous work employs modality-specific methods (gestures, spatial modification of classifier hand shapes, non-manual markers), cross-modal methods (mouthing), and modality-independent methods (such as context, and anaphora resolution) (Johnston and Schembri, 2007; Quer and Steinbach, 2015; Grimm et al., 2024).

Computational Approaches to Sign Languages. Despite progress in sign recognition (Al Abdullah et al., 2024), most work focuses on isolated signs (recognising one sign at a time) rather than a sequence of signs, covering not more than 50 signs

(Koller, 2020; Al Abdullah et al., 2024). Recent work focuses on creating new datasets: the PopSign ASL dataset (Starner et al., 2023) enables recognition of isolated signs, Neubauer (2024) identifies confusion patterns among visually similar signs (addressing homonym identification), and Ortega et al. (2025) document iconicity and concreteness norms across BSL and DGS. Grimm et al. (2024) address computational disambiguation for sign languages by using transformer-based models on the RWTH-PHOENIX-Weather Database (Koller et al., 2015), finding that approximately 64% of cases represent homonymous expressions but noting that fine-grained semantic disambiguation remains challenging. Schulder et al. (2024) develop the Multilingual Sign Language Wordnet (MSL-WN), revealing only 16% synset overlap between sign and spoken languages, confirming systematic differences exist but noting that “the nature, frequency, and distribution of these differences remain unexplored.”

Research Gap. Previous work has shown that spoken and sign languages differ in how they resolve sense disambiguation of a word and a sign. However, it remains unknown how their senses correspond. While recent work has emphasized the need for linguistically-informed sign language processing models (Yin et al., 2021), cross-modal ambiguity mapping remains largely unaddressed. Our work addresses this gap by presenting two methods to identify the type of sense correspondence between a word and its sign translation(s). Furthermore, previous datasets in sign languages are limited in scope. For instance, PopSign ASL (Starner et al., 2023) focuses on identifying homonyms in sign languages, but it is not of relevance to spoken languages. MSL-WN (Schulder et al., 2024) links sign languages to the multilingual WordNet, but it does not annotate correspondence between word uses and signs. Our dataset provides 1,404 manually annotated mappings between German word uses and DGS sign IDs across 32 words and 49 signs, with three types of sense correspondence.

3 Our Dataset

To investigate sense correspondence between German and DGS, we construct a novel dataset by combining complementary linguistic resources.

3.1 Data Sources

We combine two complementary linguistic resources: the German Word Usage Graphs (D-WUGs) (Schlechtweg et al., 2024) and the Digital Dictionary of German Sign Language (DW-DGS) (Langer et al., 2024).

D-WUG provides word uses with human-annotated semantic proximity judgments on a four-point scale (Schlechtweg et al., 2020). Uses are compiled into weighted, undirected graphs where nodes represent individual uses and edge weights correspond to median semantic proximity judgments. Correlation Clustering infers sense groupings a posteriori (Bansal et al., 2004), preserving gradedness while identifying empirically grounded semantic structures. Our analysis draws from three German datasets: DWUG_DE (10 matched words), DiscoWUG (13 matched words), and RefWUG (9 matched words), spanning two historical periods (1800–1899 and 1946–1990).

DW-DGS provides corpus-validated sign senses through video recordings and micons (moving icons) representing signs visually (Langer et al., 2024). Each unique sign receives an ID, making it easily distinguishable. The dictionary provides German translation equivalents and “Erklärung” (clarification, or the explanation of the sense).

Video IDs Instead of Glosses. Following the DGS’s practice (Otte et al., 2022), we use video IDs rather than sign glosses to mark distinct signs, as the glosses do not reliably capture homonymy and polysemy in DGS. This issue also applies to ASL (see two examples below):

Example 1: FRECH (DGS)

Homonymous Overlap. A single sign form corresponds to two completely unrelated meanings:

- Meaning 1: "frech" (cheeky/impudent) describing behavior
- Meaning 2: "USA" (West Berlin regional variant) referring to the country

Using a single gloss obscures these unrelated meanings sharing the same sign form.

Example 2: DEAF (ASL)

Movement and Location Variation. Analysis of 1,618 tokens from seven U.S. regions reveals that grammatical function is the primary constraint on this variation (Bayley et al., 2000). Three phonologically distinct variants all receive the gloss "DEAF":

- Variant A: Citation Form (ear to chin, downward path)
- Variant B: Reversed (chin to ear, upward path)
- Variant C: Contact-Cheek (cheek only, no path movement)

3.2 Human Annotation

We search DW-DGS for each unique word from the D-WUG dataset and identify 32 matching entries (split as 21/11 for "train + val + test_overlap" and test_no_overlap sets). Human annotation focuses on mapping each word use within D-WUG to its corresponding DGS sign(s), based on the German translation equivalents and "Erklärung" (clarifications) in DW-DGS. We use DW-DGS entry ID numbers instead of glosses to mark signs, as multiple DGS signs may share the same gloss but associate with different DW-DGS entries. Our annotation labels three sense correspondence types:

Type 1 (One-to-Many). A German word is polysemous or homonymous and corresponds to multiple DW-DGS signs, indicating that DGS distributes the word senses across several signs.

Type 2 (Many-to-One). Multiple German words correspond to a single polysemous or homonymous DW-DGS sign, suggesting that DW-DGS compresses several word senses into one sign form.

Type 3 (One-to-One). A German word and its DW-DGS sign translation exhibit parallel senses.

No Match. Although the German word appears in both WUG and DGS, their senses do not match.

Our human annotation proceeds as follows: (1) extract all word uses from WUG for each target word, (2) collect all DW-DGS signs matching the word with their German translations and sense clarifications, (3) manually map each word use to appropriate sign ID(s), (4) manually assign a suitable correspondence type. For example, "Behandlung" (treatment/handling) maps to two signs (IDs 637, 999) for medical versus processing contexts, yielding Type 1; "Abend" (evening) maps to one sign (ID 19) that also corresponds to "Nacht" (night), yielding Type 2.

Split	Level	T1	T2	T3	NM
train + val + test_overlap	mapping	573	147	236	17
test_no_overlap	mapping	168	84	87	92
train + val + test_overlap	word	6	6	7	2
test_no_overlap	word	3	3	3	2
Total: 1,404 mappings from 32 words and 49 signs					

Table 1: Dataset statistics showing balanced distribution across sense correspondence types, with Type 1 (one-to-many) being most common at the instance level (573 instances, 59.0% of development set).

Details of our dataset are outlined in Table 1. In the development set (train+val+test_overlap), type distribution of headwords has relatively balanced coverage: Type 1 (28.6%), Type 2 (28.6%), Type 3 (33.3%), and No Match (9.5%).

4 Our Approach

We implement two computational methods to automatically identify sense correspondence types between German words and their DGS sign translations.

Exact Match (EM) retrieves candidate DW-DGS signs by looking for lexical overlap between the two D-WUG entries (a headword and its word use) and the two DW-DGS entries (German translations and "Erklärung" of each sign). When a match is found, the corresponding DW-DGS video entry ID is retrieved as a candidate. All candidates are then ranked by computing the semantic similarities between D-WUG and DW-DGS entries based on their embeddings. Note that EM uses semantic similarity only for ranking retrieved candidates, whereas SS uses it for both retrieval and ranking.

Semantic Similarity (SS) encodes the same D-WUG entries and the same DW-DGS entries (as in EM) into embeddings by using SBERT (Reimers and Gurevych, 2019b). We use q to denote the concatenated embedding of the D-WUG entries, while using m to denote the concatenated embedding of the DW-DGS entries. We compute the cosine similarity between q and m:

$$\text{sim}(q, m) = \frac{\mathbf{e}_q \cdot \mathbf{e}_m}{|\mathbf{e}_q||\mathbf{e}_m|} = \cos(\theta) \tag{1}$$

The similarity score ranges from -1 to 1, with values closer to 1 indicating stronger semantic alignment. We evaluate four sentence embedding models: paraphrase-multilingual-MiniLM-L12-v2, all-MiniLM-L6-v2, German_Semantic_STS_V2, and German-roberta-sentence-transformer-v2 (Reimers and Gurevych, 2019a, 2020).

Sense correspondence type. For each German word w, we collect the predicted DGS sign IDs, together with their semantic similarity scores that are previously denoted by $\text{sim}(q, m)$. We let V_w be a set of DGS sign IDs predicted for word w. The sense Correspondence Type $\text{T}(w)$ is then assigned according to the following rules:

$$\text{T}(w) = \begin{cases} \text{No Match}, & \text{if } |V_w| = 0, \\ \text{Type } 1, & \text{if } |V_w| > 1, \\ \text{Type } 2\text{ }, & \text{if } |V_w| = 1 \text{ AND sim} < \tau, \\ \text{Type } 3, & \text{if } |V_w| = 1 \text{ AND sim} \geq \tau \end{cases} \tag{2}$$

This applies to both SS and EM methods.

5 Experimental Setup

We evaluate our methods across multiple configurations to assess their effectiveness in identifying sense correspondence types.

5.1 Data Splits

For the development set, we partition the manually annotated data into three splits: (i) **the train split** (723 mappings), which serves as the candidate pool for retrieving potential sign matches; (ii) **the validation split** (100 mappings), used for hyperparameter tuning via grid search; and (iii) **the test_overlap split** (150 mappings), used for model comparison and evaluation.

5.2 Evaluation Setup

We conduct experiments in two setups: (i) **With vocabulary overlap**: the test_overlap split contains words and signs that are present in the train set, enabling direct comparison between EM and SS methods and (ii) **Without vocabulary overlap**: the test_no_overlap set contains entirely different words and signs not in the train set, measuring whether SS can generalize to novel vocabulary. Since EM requires lexical matches, only SS is evaluated in the zero-overlap scenario.

5.3 Hyperparameter optimization

The SS method has two hyperparameters: (i) similarity threshold τ, determining minimum cosine similarity required and (ii) top-k, controlling how many high-scoring candidates are selected from each data source before merging. We optimise these hyperparameters via grid search on the validation split, totalling 12 configurations: $\tau \in \{0.65, 0.70, 0.75, 0.80\}$ and $k \in \{3, 5, 7\}$. Grid search is conducted separately for each embedding model.

5.4 Evaluation Metrics

Our metric is accuracy, defined as the proportion of cases where top-ranked predicted signs (including ties) match our human annotation. For No Match cases, a prediction is considered correct only if our methods return no prediction, i.e., no candidate exceeds the similarity threshold. Additionally, we evaluate the ranking quality beyond the top prediction by reporting Precision@K, which measures whether the correct sign appears within the top K ranked candidates, where $K \in \{1, 3, 5, 10\}$ (Järvelin and Kekäläinen, 2002; Manning et al., 2008).

Lastly, we break down our evaluation into individual sense correspondence types, examining performance gap across the three correspondence types (one-to-many, many-to-one, and one-to-one) as well as No Match cases. Additionally, we conduct an error analysis reporting error rate per type, and then look into the prediction agreement between EM and SS, for instance, analysing how often both methods succeed, how often both fail, how often only one method succeeds.

5.5 Ablation Setups

We experiment with two ablation setups to evaluate the impact of different input components:

- D-WUG entries: (i) **Full Context**, which combines a German word and its word use; (ii) **Word Only**, which uses the word only; and (iii) **Sentence Only**, which uses the word use only.

- DW-DGS entries: (i) **Base**, which uses both German translation equivalents (GT) and “Erklärung”, and (ii) **GT Only**, which uses German translations only.

Model	Thr.	K	Acc.	Imp.
paraphrase-multi-MiniLM-L12-v2	0.65	3	78.57	+8.33
all-MiniLM-L6-v2	0.70	3	73.81	-1.19
German_semantic_sts_v2	0.80	3	72.62	+4.76
German-roberta-sent-v2	0.80	3	69.05	-1.19

Table 2: Optimal hyperparameters for each embedding model on the validation split. The paraphrase-multilingual model achieves the highest accuracy (78.57%) and shows the largest improvement over exact matching (+8.33 pp).

Model	EM	SS	Imp.
all-MiniLM-L6-v2	71.31	88.52	17.21
German_semantic_sts_v2	70.49	87.70	17.21
paraphrase-multi-MiniLM-L12-v2	68.85	86.89	18.03
German-roberta-sent-v2	71.31	84.43	13.11

Table 3: Model performance on test_overlap split demonstrates that all embedding models substantially outperform exact matching, with all-MiniLM-L6-v2 achieving the best accuracy (88.52%, +17.21 pp improvement).

6 Results

We present the performance of our methods across different evaluation scenarios and correspondence types.

6.1 Model Selection

Table 2 reports the hyperparameter configuration of each model. All the hyperparameters are tuned by using grid search on the validation data split. Then, we evaluate the models with these hyperparameters on the test_overlap set. We find that paraphrase-multi-MiniLM-L12-v2 achieves the best accuracy (78.57%) on the validation split, while all-MiniLM-L6-v2 achieves the best accuracy (88.52%) on the test_overlap split, with SS outperforming EM by a 17.21 percentage point. Thus, we use all-MiniLM-L6-v2 for the remaining analyses.

6.2 Overall Results

Table 4 presents model accuracies on the test_overlap split. We see that SS outperforms EM in all cases by a 17.21 percentage point, indicating approximately 24% relative improvement. Both SS and EM achieve a perfect score (100%) in “No Match” cases. Thus, the improvement of SS stems from “Match” cases, where SS achieves 86.67% accuracy compared to EM’s 66.67% (+20.0 pp).

Category	EM	SS	Imp.
Overall (n=122)	71.31	88.52	17.21
Match (n=105)	66.67	86.67	20.00
No Match (n=17)	100.0	100.0	0.0

Table 4: Overall performance comparison shows semantic similarity (SS) outperforms exact matching (EM) by 17.21 percentage points, with the improvement stemming entirely from ’Match’ cases where sense correspondence exists.

Type	n	EM	SS	Imp.
Type 1	48	41.7	93.8	52.1
Type 2	21	66.7	66.7	0.0
Type 3	36	100.0	88.9	-11.1
No Match	17	100.0	100.0	0.0

Table 5: Type-specific accuracy reveals dramatic differences: SS excels at Type 1 (one-to-many) with +52.1 pp improvement, EM performs best for Type 3 (one-to-one), while Type 2 (many-to-one) remains equally challenging for both methods.

6.3 Type-Specific Results

Table 5 shows how model accuracies vary across different sense correspondence types. Type 1 seems most challenging for EM (41.7%) but benefits dramatically from SS (93.8%, +52.1 pp), demonstrating that SS relying on SBERT can identify one-to-many sense correspondence type in almost all cases. For Type 3, EM achieves a perfect score EM performance (100%), while SS lags behind (88.9%, -11.1 pp). This suggests that when a German word and its sign translation has parallel senses, our SS is not so reliable, which may introduce noise and produce wrong matches. However, the advantage of EM over SS is likely due to a dataset artefact, namely the vocabulary overlap between data splits, which may not be generalisable to other datasets. For Type 2, both approaches are on par, indicating equal challenges for both. For No Match, both approaches achieve a perfect score (100%).

6.4 Ranking Quality and Error Patterns

In Table 6, we find that SS improves from P@1 (79.5%) to P@3 (88.5%), while EM from 71.3% to 91.8%. Both plateau at K=3. In Table 7, we see that SS successfully predicts 20.5% of cases in which EM fails (25 instances), while it fails in 3.3% of cases where EM succeeds (4 instances), yielding a net gain of +17.2 pp. In 68.0% of cases, both methods succeed, whereas both fail in 8.2% of cases (10 instances).

Method	Category	n	P@1	P@3
SS	Overall	122	79.5	88.5
	Match	105	76.2	86.7
	No Match	17	100.0	100.0
EM	Overall	122	71.3	91.8
	Match	105	66.7	90.5
	No Match	17	100.0	100.0

Table 6: Ranking quality analysis shows both methods plateau at P@3, with EM achieving slightly higher precision (91.8%) than SS (88.5%) when considering top-3 predictions, suggesting EM provides better candidate ranking despite lower top-1 accuracy.

Pattern	Count	%
Both Succeed	83	68.0
SS Success, EM Fail	25	20.5
EM Success, SS Fail	4	3.3
Both Fail	10	8.2

Table 7: Error pattern analysis reveals SS succeeds in 20.5% of cases where EM fails, while failing in only 3.3% of cases where EM succeeds, yielding a net gain of +17.2 pp and demonstrating complementary strengths.

6.5 Confusion Matrix

We use a confusion matrix to report the agreement between ground-truth and model predictions (by SS) of sense correspondence types on the test_overlap set (20 words) (see Tables 14 and 8). Overall agreement reaches 60% (12/20 words with correct predictions of sense correspondence types). No Match achieves perfect agreement (100%, 2/2), followed by Type 3 at 85.7% (6/7). Type 1 achieves 66.7% agreement (4/6). Our SS fails to identify Type 2 instances (0/5), never predicting Type 2, instead misclassifying those cases as No Match (2 words) or Type 3 (3 words).

6.6 Ablation Studies

Table 9 reports ablation results for both data sources. For D-WUG entries, “Word Only” achieves the best SS accuracy (91.80%), outperforming “Full Context” by +3.28 pp, while “Sentence Only” yields the best EM performance (73.77%). The underperformance of “Full Context” suggests that jointly encoding the word with its usage introduces noise that negatively affects predictions. For DW-DGS entries, including “Erklärungen” (dictionary definitions) yields no performance difference compared to using German translations alone for both methods, indicating that translation equivalents contain sufficient semantic information.

Type	Total	Correct	Agr. (%)
No Match	2	2	100.0
Type 1	6	4	66.7
Type 2	5	0	0.0
Type 3	7	6	85.7
Overall	20	12	60.0

Table 8: Agreement rates between ground truth and predictions show perfect performance for No Match (100%), strong performance for Type 3 (85.7%), moderate for Type 1 (66.7%), but complete failure for Type 2 (0%), resulting in 60% overall agreement.

D-WUG Entries Ablation			
Input Mode	**EM**	**SS**	**Imp.**
Word Only	72.95	91.80	18.85
Full Context	71.31	88.52	17.21
Sentence Only	73.77	88.52	14.75
DW-DGS Entries Ablation			
Configuration	**EM**	**SS**	**Imp.**
Base (GT + “Erklärung”)	71.31	88.52	17.21
GT Only	71.31	88.52	17.21
Difference	0.00	0.00	0.00

Table 9: Ablation studies show that (1) encoding the German word alone achieves the best SS accuracy (91.80%), outperforming full context by +3.28 pp, suggesting word usage context introduces noise, and (2) including dictionary definitions provides no gain over German translations alone (all values in %).

6.7 Generalization (test_no_overlap)

Our method strongly relies on vocabulary overlap between data splits. In Table 10, we find that model performance drops sharply from the overlap to the no-overlap setting (88.52% to 17.59%), suggesting that the overlap between the test_overlap split and the train split is crucial. Without such overlap, our SS based on word uses only cannot address sense correspondence between a word and its sign translation.

Our ablation studies support this finding: Using “Word Only” performs better than using “Full Context” that combines word and its uses (91.80% vs. 88.52%) as shown on Table 9, indicating the importance of vocabulary. Although incorporating word uses from WUG datasets does not improve performance here, such contextual information may be beneficial for other tasks across spoken and sign languages.

Metric	W/Ovlp	W/o Ovlp	Gap
Overall Acc. (%)	88.52	17.59	-70.93
Match (n)	105	431	—
Match Acc. (%)	86.67	0.00	-86.67
No Match (n)	17	92	—
No Match Acc. (%)	100.00	100.00	0.00
Type 1 Agr. (%)	66.7	0.0	-66.7
Type 3 Agr. (%)	85.7	0.0	-85.7
Overall Agr. (%)	60.0	18.2	-41.8

Table 10: test_no_overlap analysis reveals severe performance degradation without vocabulary overlap (88.52% to 17.59%, -70.93 pp), demonstrating that the model's success critically depends on seeing the same words during training rather than generalizing to semantic patterns.

6.8 Parts-of-speech Analysis

We also conduct a parts-of-speech analysis to examine whether different word classes exhibit distinct sense correspondence patterns. Appendix C provides detailed results. Table 15 shows that verbs predominantly map to Type 1 (3 of 6), nouns favor Types 2 and 3 combined (8 of 10), and adjectives split evenly between Types 1 and 3. Table 16 reveals that in the zero-overlap scenario, our SS model predicts "No Match" for 9 of 11 words despite only 2 being genuine "No Match" cases, further confirming poor test_no_overlap without vocabulary overlap.

7 Discussion

Our findings reveal systematic patterns in how senses are organized across spoken and sign language modalities. Our analysis investigates how senses are organised differently across spoken and sign language modalities. The relatively balanced distribution of headwords across sense correspondence types (Type 1: 28.6%, Type 2: 28.6%, Type 3: 33.3%) demonstrates that no single correspondence type dominates sense organisation across spoken and sign languages, while confirming the differences between the two modalities (Schulder et al., 2024; Taub, 2001).

Sense correspondence between spoken and sign languages is dominated by Type 1 (50.0%, 13 out of 26 instances) at the level of word–to–sign pairs, where the senses of a polysemous or homonymous spoken word are distributed across multiple signs. This aligns with findings by Kristoffersen and Troelsgård (2010) that sign language uses the visual-gestural modality to encode fine-grained senses across multiple signs.

Performance gaps across correspondence types highlight the limitations of our methods. For Type 1 cases, SS outperforms EM, indicating that the SBERT that SS relies on can address correspondence between a polysemous word and multiple sign candidates, precisely the scenario where EM fails. For Type 3, where words and signs exhibit parallel senses, SBERT-based SS can sometimes overgeneralise by matching semantically similar but incorrect sign candidates. Type 2 is equally challenging for both EM and SS, as both methods struggle to distinguish whether a single sign has one sense or multiple senses.

Our approaches are evaluated using both accuracy (accounting for ties) and P@K. The 9 percentage point difference between accuracy (88.52%; see Table 3) and P@1 (79.5%; see Table 6) indicates that in approximately 9% of cases on the test_overlap set, SS identifies the ground-truth sign but does not rank it first. Our analysis also shows that SS is effective at retrieving semantically related correspondence that EM fails to detect, while EM provides better ranking performance than SS when the correct candidates are retrieved.

8 Conclusion

This work provides the first structured analysis of sense correspondence types between spoken German and German Sign Language. Our work analyses how lexical senses align across spoken and sign languages. We manually annotated sense correspondence between German words and DGS signs, and presented computational methods to identify the types of sense correspondence. We found that the senses of German words and their sign translations are organised differently across three correspondence types.

We contribute a dataset of 1,404 manually annotated word use–to–sign ID mappings derived from 32 words 49 signs, establishing the first resource of its own kind. Our computational evaluation identifies which correspondence patterns are identifiable: SS is effective when a German word corresponds to multiple signs (+52.1 pp), EM performs best when a word and its sign translation have parallel senses, while both methods struggle when multiple German words correspond to a single sign. These findings reveal which correspondence patterns current approaches can identify: Type 1 (one-to-many) correspondences are highly identifiable through

SS, Type 3 (one-to-one) benefits from EM, while Type 2 (many-to-one) remains challenging for both methods. Our findings advance prior work. While Schulder et al. (2024) demonstrate systematic differences exist through 16% synset overlap, noting the nature, and distribution of these differences remain unexplored, our analysis clarifies how these differences manifest across sense correspondence types and which cases are more difficult to identify using our approaches. While fine-grained sign language sense disambiguation remains challenging (Grimm et al., 2024), our results identify specific scenarios in which our methods succeed and others where they fall short.

Future work should enlarge our dataset, expand to other languages, include annotations conducted by native German–DGS bilingual speakers, explore multimodal embeddings that integrate sign videos, investigate dialectal variation using DW-DGS, and study semantic change across spoken and sign languages.

Limitations

Our annotations were not conducted by native German–DGS bilingual speakers. We restricted each word–sign pair to a single sense correspondence type. Our dataset only contains 32 words and 49 signs, which is small; however, our focus is 1,404 human-annotated mappings from word usages to signs. Our findings are limited to the German-DGS language pair and may not generalise to other languages. Finally, our approach relies heavily on vocabulary overlap; without such overlap, accuracy drops largely from 88.52% to 17.59%.

References

Bashaer A. Al Abdullah, Ghada A. Amoudi, and Hanan S. Alghamdi. 2024. Advancements in sign language recognition: A comprehensive review and future prospects. *IEEE Access*, 12:128871–128895.

B. Bahan and J. Dannis. 1996. *Come Sign With Us: Sign Language Activities for Children*, 2 edition. Gallaudet University Press, Washington, DC.

Nikhil Bansal, Avrim Blum, and Shuchi Chawla. 2004. Correlation clustering. *Machine Learning*, 56(1):89–113.

Robert Bayley, Ceil Lucas, and Mary Rose. 2000. Phonological variation in American Sign Language: The case of 1 handshape. *Language Variation and Change*, 12(1):25–48.

V. Fromkin, R. Rodman, and N. Hyams. 2010. *An Introduction to Language*. Cengage Learning.

Jana Grimm, Miriam Winkler, Oliver Kraus, and Tanalp Agustoslu. 2024. Sign language sense disambiguation. *Preprint*, arXiv:2409.08780.

I. G. Gwammaja. 2025. A semantic description of homosigns in hausa sign language. *LASU Postgraduate School Journal (LPSJ)*, 2(2):510–527.

Janosch Haber and Massimo Poesio. 2024. Polysemy—evidence from linguistics, behavioral science, and contextualized language models. *Computational Linguistics*, 50(1):351–417.

J.R. Hurford, B. Heasley, and M.B. Smith. 2007. *Semantics: A Coursebook*. Cambridge University Press.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Trevor Johnston and Adam Schembri. 2007. *Australian Sign Language (Auslan): An Introduction to Sign Language Linguistics*. Cambridge University Press, Cambridge, UK.

Ekaterini Klepousniotou. 2002. The processing of lexical ambiguity: Homonymy and polysemy in the mental lexicon. *Brain and Language*, 81(1):205–223.

Oscar Koller. 2020. Quantitative survey of the state of the art in sign language recognition. *Preprint*, arXiv:2008.09918.

Oscar Koller, Jens Forster, and Hermann Ney. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125.

Jette Hedegaard Kristoffersen and Thomas Troelsgård. 2010. Making a dictionary without words: Lemmatization problems in a sign language dictionary. In *eLexicography in the 21st Century: New Challenges, New Applications. Proceedings of eLex 2009, Louvain-la-Neuve, 22-24 October 2009*, volume 7 of *Cahiers Du Cental*, pages 165–172, Louvain. Presses Universitaires de Louvain.

Gabriele Langer, Anke Müller, Sabrina Wähl, Felicitas Otte, Lea Sepke, and Thomas Hanke. 2024. Introducing the DW-DGS – the digital dictionary of DGS. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 194–203, Torino, Italia. ELRA and ICCL.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

William Neubauer. 2024. PopSign ASL V2.0: A large isolated sign language dataset. Bachelor's thesis, Georgia Institute of Technology, Atlanta, GA, December.

Gerardo Ortega, Annika Schiefner, Nia Lazarus, and Pamela Perniss. 2025. A lexical database of British Sign Language (BSL) and German Sign Language (DGS): Iconicity ratings, iconic strategies, and concreteness norms. *Behavior Research Methods*, 57(5):139.

Felicitas Otte, Anke Müller, Gabriele Langer, Sabrina Wähl, and Thomas Hanke. 2022. Sign representation in the dw-dgs. Technical report, Project Note AP11-2021-01, Universität Hamburg, DGS-Korpus project, IDGS ....

Roland Pfau and Markus Steinbach. 2016. Modality and meaning: Plurality of relations in german sign language. *Lingua*, 170:69–91.

Josep Quer and Markus Steinbach. 2015. Ambiguities in sign languages. *The Linguistic Review*, 32.

K.M. Tahsin Hassan Rahit, Khandaker Tabin Hasan, Md. Al Amin, and Zahiduddin Ahmed. 2018. BanglaNet: Towards a WordNet for Bengali language. In *Proceedings of the 9th Global Wordnet Conference*, pages 1–9, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.

Nils Reimers and Iryna Gurevych. 2019a. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Nils Reimers and Iryna Gurevych. 2019b. Sentence-transformers: Multilingual sentence, paragraph, and image embeddings using bert & co. Retrieved November 22, 2025.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

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, Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Nikolay Arefyev. 2024. Sense through time: Diachronic word sense annotations for word sense induction and lexical semantic change detection. *Language Resources and Evaluation*.

Marc Schulder, Sam Bigeard, Maria Kopf, Thomas Hanke, Anna Kuder, Joanna Wójcicka, Johanna Mesch, Thomas Björkstrand, Anna Vacalopoulou, Kyriaki Vasilaki, Theodore Goulas, Stavroula-Evita Fotinea, and Eleni Efthimiou. 2024. Signs and synonymity: Continuing development of the multilingual sign language Wordnet. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 343–353, Torino, Italia. ELRA and ICCL.

Masda Surti Simatupang. 2007. How ambiguous is the structural ambiguity. *Lingua Cultura*, 1(2):99–104.

Thad Starner, Sean Forbes, Matthew So, David Martin, Rohit Sridhar, Gururaj Deshpande, Sam Sepah, Sahir Shahryar, Khushi Bhardwaj, Tyler Kwok, Daksh Sehgal, Saad Hassan, Bill Neubauer, Sofia Anandi Vempala, Alec Tan, Jocelyn Heath, Unnathi Utpal Kumar, Priyanka Vijayaraghavan Mosur, Tavenner M. Hall, and 5 others. 2023. Popsign asl v1.0: an isolated american sign language dataset collected via smartphones. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Sarah Taub. 2001. Language from the body: Iconicity and metaphor in american sign language.

Hongzhi Xu, Jingxia Lin, Sameer Pradhan, Mitchell Marcus, and Ming Liu. 2024. Annotating Chinese word senses with English WordNet: A practice on OntoNotes Chinese sense inventories. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1187–1196, Torino, Italia. ELRA and ICCL.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. 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)*, pages 7347–7360, Online. Association for Computational Linguistics.

APPENDIX

A More Tables

Table 11: Optimal hyperparameter configurations for each model (complete version).

Model	Thr.	K	Acc.	Imp.
paraphrase-multi-lingual-MiniLM-L12-v2	0.65	3	78.57	+8.33
all-MiniLM-L6-v2	0.70	3	73.81	-1.19
German_Semantic_STS_V2	0.80	3	72.62	+4.76
german-roberta-sentence-trans-former-v2	0.80	3	69.05	-1.19

Models ranked by optimization semantic accuracy. Thr.=Threshold, K=Top-K, Acc.=Semantic Accuracy (%), Imp.=Improvement (%). Batch size=64 for all models. Improvement refers to SS performance over EM on the validation set.

Table 12: Precision@K results on test split (complete version).

Meth.	Category	n	P@1	P@3	P@5
SS	Overall	122	79.5	88.5	88.5
	Match	105	76.2	86.7	86.7
	No-Match	17	100.0	100.0	100.0
EM	Overall	122	71.3	91.8	91.8
	Match	105	66.7	90.5	90.5
	No-Match	17	100.0	100.0	100.0

P@K = percentage of cases where correct sign appears within top K predictions. P@10 values identical to P@5 (omitted for space). For No Match cases, P@K measures correct abstention.

Table 13: Input component ablation results on test split (complete version).

Input Mode	EM	SS	Imp.	Conf.
Word Only	72.95	91.80	18.85	0.830
Full Context	71.31	88.52	17.21	0.746
Sentence Only	73.77	88.52	14.75	0.746

EM=EM Accuracy (%), SS=Semantic Accuracy (%), Imp.=Improvement (%), Conf.=Average Confidence. All configurations use best model (all-MiniLM-L6-v2) with optimized hyperparameters (threshold=0.70, top_k=3). Full Context represents baseline configuration.

GT ↓ / Pred →	NM	T1	T3	Total
No Match	2	0	0	2
Type 1	1	4	1	6
Type 2	2	0	3	5
Type 3	1	0	6	7
Total	6	4	10	20

Table 14: Confusion matrix on test set (20 words) shows the model never predicts Type 2, instead misclassifying all 5 Type 2 cases as either No Match (2) or Type 3 (3), indicating fundamental difficulty distinguishing polysemous signs.

B Computational Typology Classification Visualization

Figure 1 shows agreement between computational typology discovery and human annotations across 20 test words. The model achieves strong agreement for No Match (100%, 2/2) and Type 3 (85.7%, 6/7), moderate agreement for Type 1 (66.7%, 4/6), but completely fails to identify Type 2 (0/5). The model never predicts Type 2, instead misclassifying these cases as No Match (2) or Type 3 (3), suggesting fundamental difficulty in distinguishing whether a single sign represents one coherent meaning or multiple unrelated meanings.

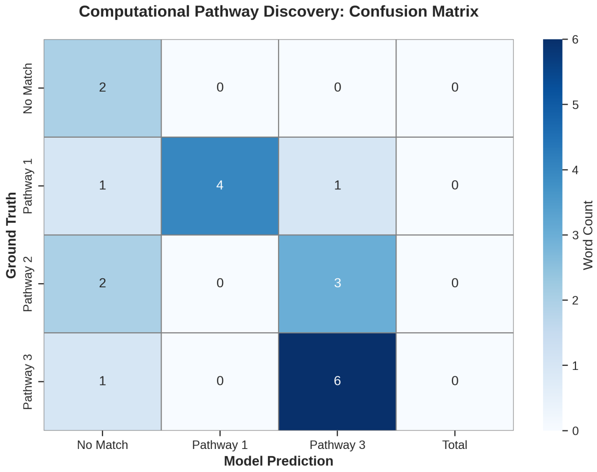

Figure 1: Confusion matrix visualizing agreement between model predictions and human annotations across 20 test words, clearly showing the model's complete inability to identify Type 2 (many-to-one) correspondence, with all 5 Type 2 cases misclassified as either No Match or Type 3.

C Parts-of-Speech Analysis

Table 15 presents the distribution of sense correspondence types across different parts of speech in the development set, revealing systematic patterns in how word classes map to correspondence types.

POS	Word	GT Type
Verb	ausbilden	Type 2
Verb	bemerken	Type 3
Verb	eintreten	Type 3
Verb	erlauben	Type 1
Verb	freigelassen	Type 1
Verb	helfen	Type 1
Noun	Museum	Type 2
Noun	Mauer	Type 1
Noun	Vorbereitung	Type 2
Noun	Westen	Type 3
Noun	Behandlung	Type 1
Noun	Entscheidung	Type 3
Noun	Frechheit	Type 2
Noun	Mut	Type 3
Noun	Seminar	Type 3
Noun	Tier	Type 2
Adjective	englisch	Type 1
Adjective	finnisch	Type 3

Table 15: Parts-of-speech distribution in development set shows verbs predominantly map to Type 1 (one-to-many, 3 of 6), nouns favor Types 2 and 3 combined (8 of 10), and adjectives split evenly, suggesting word class influences sense correspondence patterns.

Table 16 compares ground-truth and model predictions in the zero-vocabulary-overlap scenario, demonstrating the model's failure to generalize beyond memorized vocabulary.

POS	Word	GT Type	Model Type
Noun	Anstellung	Type 1	No Match
Adjective	billig	Type 3	No Match
Noun	Zufall	Type 3	No Match
Noun	Presse	Type 3	No Match
Verb	packen	Type 1	No Match
Verb	anpflanzen	No Match	No Match
Verb	niederschlagen	No Match	No Match
Verb	abbauen	Type 2	Type 3
Verb	artikulieren	Type 3	No Match
Noun	Schmiere	Type 2	No Match
Noun	Titel	Type 2	No Match

Table 16: Model predictions in zero-vocabulary-overlap scenario show the model incorrectly predicts "No Match" for 9 of 11 words despite only 2 genuine cases, confirming poor test_no_overlap and over-reliance on lexical memorization.