SenWiCh: Sense-Annotation of Low-Resource Languages for WiC using Hybrid Methods

Roksana Goworek[1,2], Harpal Karlcut[1], Hamza Shezad[1], Nijaguna Darshana[1], Abhishek Mane[1], Syam Bondada[1], Raghav Sikka[1], Ulvi Mammadov [1], Rauf Allahverdiyev[1], Sriram Purighella[1], Paridhi Gupta[1], Muhinyia Ndegwa[1], Bao Khanh Tran[1], Haim Dubossarsky[1,2,3]

[1]Queen Mary University of London, [2]The Alan Turing Institute, [3]University of Cambridge,

Abstract

This paper addresses the critical need for high-quality evaluation datasets in low-resource languages to advance cross-lingual transfer. While cross-lingual transfer offers a key strategy for leveraging multilingual pretraining to expand language technologies to understudied and typologically diverse languages, its effectiveness is dependent on quality and suitable benchmarks. We release new sense-annotated datasets of sentences containing polysemous words, spanning ten low-resource languages across diverse language families and scripts. To facilitate dataset creation, the paper presents a demonstrably beneficial semi-automatic annotation method. The utility of the datasets is demonstrated through Word-in-Context (WiC) formatted experiments that evaluate transfer on these low-resource languages. Results highlight the importance of targeted dataset creation and evaluation for effective polysemy disambiguation in low-resource settings and transfer studies. The released datasets and code aim to support further research into fair, robust, and truly multilingual NLP.

1 Introduction

Cross-lingual transfer is a key strategy in modern NLP, particularly for low-resource languages, where training data is scarce. By leveraging multilingual pretraining, models can transfer task-specific abilities from high-resource languages to low-resource ones, expanding access to language technologies for underrepresented communities (He et al., 2021; Ponti et al., 2018; Wei et al., 2021).

Despite its promise, transfer learning is not universally effective across tasks or languages. Studies on tasks like POS tagging, NER, NLI, QA, and sentiment analysis (Pires et al., 2019; Dolicki and Spanakis, 2021; Srinivasan et al., 2021; Lauscher et al., 2020; Ahuja et al., 2023), as well as polysemy disambiguation (Raganato et al., 2020; Dairkee and Dubossarsky, 2024), show that cross-lingual transfer can be inconsistent and, in some cases, fail entirely. This is also true for generative models (Robinson et al., 2023; Shaham et al., 2024; Chirkova and Nikoulina, 2024), with particularly poor performance in low-resource languages, highlighting the need for more robust and language-inclusive transfer.

A main obstacle for transfer is the lack of high-quality datasets in low-resource and typologically diverse languages. Without these benchmarks, assessing transfer performance, let alone training models on target languages, remains a formidable challenge. This lacking is largely due to the scarcity of linguistic resources in low-resource languages. For instance, Wiktionary contains over a million entries for German, English, French, Chinese, and Russian, but fewer than 100,000 for Punjabi and Marathi (Wikimedia Foundation, 2025).

This lack of resources underscores the urgent need for dedicated datasets to evaluate and refine transfer techniques for underrepresented languages, which this work addresses by developing a semi-automatic method for sense annotation in polysemy and generating resources in ten languages.

We focus on the task of polysemy disambiguation, as it particularly challenges cross-lingual transfer by revealing structural and semantic differences between languages. While some NLP tasks, like sentiment analysis, rely on meaning preservation across languages, where direct translation can maintain performance, polysemy is highly language-specific (Rzymski et al., 2020), making it a rigorous test of a model's ability to generalize across languages. For example, the English word "movement" refers to both physical motion and a political or social movement. However, its Polish translation, "ruch", also encompasses these two meanings, but additionally means "traffic", a sense not covered by the English word. Conversely, "movement" in English can also refer to a section of a musical composition.

Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2025), pages 61–74
August 1, 2025 ©2025 Association for Computational Linguistics

Polysemy disambiguation has long been considered a hallmark of human cognition and a central challenge in NLP (Navigli, 2009; Bevilacqua et al., 2021). A model that can accurately distinguish between different senses of a word must capture linguistic subtleties, metaphorical meanings, and even emerging word usages, much like human speakers. Thus, success in cross-lingual polysemy disambiguation would suggest a model's ability to generalize deep semantic understanding, beyond surface-level patterns in a single language. While many high-resource languages already benefit from sense-annotated datasets (see §2), low-resource languages remain largely unrepresented in this area. Existing contextualized models can process polysemous words within downstream tasks (Loureiro et al., 2021; Ushio et al., 2021), but sense disambiguation remains a major challenge across dozens of languages (Pilehvar and Camacho-Collados, 2019a; Raganato et al., 2020; Martelli et al., 2021; Liu et al., 2021).

Beyond NLP, polysemy also presents difficulties in multimodal models, such as object detection systems, where the same word can refer to multiple visual categories (Calabrese et al., 2020). This suggests that solving polysemy is not just beneficial for language tasks but has broader implications for AI reasoning and multimodal understanding.

Our Contributions Despite extensive work on polysemy disambiguation in high-resource languages, datasets for low-resource languages remain scarce. We address this gap with the following contributions:

- **Sense-annotated datasets:** We release both WSD-style sense-annotated corpora and WiC-style evaluation datasets for ten low-resource languages.[1] The WiC format supports direct comparison with existing experiments in other languages, enabling strong cross-lingual baselines.

- **Annotation tool:** To facilitate further resource development, we release a hybrid semi-automated annotation tool.[2]

Together, these contributions represent a crucial step toward advancing fair, robust, and truly multilingual NLP by enabling evaluation and development in languages that have been largely neglected.

2 Related Work

2.1 Transfer Studies

Zero-shot cross-lingual transfer has been widely studied, with mixed findings on its effectiveness, particularly in polysemy disambiguation. While some studies highlight transfer potential across languages, others expose significant limitations, especially in tasks that depend on fine-grained semantic distinctions.

Lauscher et al. (2020) examined zero-shot transfer performance across 17 languages and five NLP tasks (excluding polysemy), evaluating XLM-R (Conneau et al., 2020) and mBERT (AI, 2018). They found that zero-shot performance drops significantly compared to full-shot settings and that transfer success correlates with factors like pretraining corpus size and linguistic similarity. These findings suggest that cross-lingual transfer is far from universal and is highly dependent on language resources and pretraining coverage.

Focusing specifically on polysemy disambiguation, Raganato et al. (2020) conducted the first large-scale cross-lingual transfer study for this task, training a model on English and evaluating on 12 other languages. While they observed some zero-shot transferability, models trained on English underperformed models trained on the target language by 10-20% when tested on German, French, and Italian, indicating that polysemy disambiguation remains language-sensitive and benefits from in-language supervision.

In contrast, Dairkee and Dubossarsky (2024) challenged the feasibility of cross-lingual transfer for polysemy disambiguation altogether. Studying English and Hindi, they found a complete lack of zero-shot transfer, suggesting that word sense distinctions may be too language-specific for direct transfer without explicit in-language supervision.

These conflicting results emphasize the need for more comprehensive transfer studies in polysemy disambiguation, particularly in low-resource languages where transfer learning is often the only viable approach due to the lack of labeled data. However, without high-quality evaluation datasets in these languages, assessing and improving transfer learning for polysemy remains an open challenge.

2.2 Polysemy Disambiguation

Word Sense Disambiguation (WSD) datasets are sense-annotated corpora consisting of sentences containing polysemous words, labeled according

[1] available at DOI: 10.5281/zenodo.15493005

[2] available at github.com/roksanagow/projecting_sentences

to their contextual meanings. WSD is inherently complex, as words vary in the number of possible senses, and the list of words differs across languages. To address this, Pilehvar and Camacho-Collados (2019a) introduced the Word in Context (WiC) formulation, which reformulated the original WSD problem, which was a multi-class classification task, into a binary classification one. Instead of assigning specific sense labels, WiC pairs two sentences containing the same word and labels them 1 (same) or 0 (different). For example:

> A **bat** flew out of the cave as the sun set.
> He swung the **bat** with all his strength.

This approach enables models to be trained directly on polysemy disambiguation by adjusting embeddings so that words with the same sense cluster together, while those with different senses are pushed apart in the resulting embedding space.

2.3 Existing Datasets

2.3.1 WSD Datasets

Word Sense Disambiguation (WSD) research has been supported by several key sense-annotated corpora and lexical resources:

SemCor (Miller et al., 1993) is a foundational English corpus containing over 226,000 sense annotations across 352 documents.

OntoNotes (Hovy et al., 2006) offers a multi-genre corpus with extensive annotations, including word senses linked to a refined sense inventory for English, Chinese and Arabic.

Senseval/SemEval Datasets have been instrumental in standardizing WSD evaluation. Notably, **Senseval-2** (Edmonds and Cotton, 2001) and **SemEval-2007 Task 17** (Pradhan et al., 2007) provided all-words WSD tasks, challenging systems to disambiguate every content word in given texts. These competitions have included data in multiple languages, such as English, Chinese, Basque, and others (Navigli et al., 2013).

CoarseWSD-20 (Loureiro et al., 2021) is a coarse-grained sense disambiguation dataset derived from Wikipedia, focusing on 20 ambiguous nouns, each with 2 to 5 senses, all in English.

FEWS (Few-shot Examples of Word Senses) (Blevins et al., 2021) addresses the challenge of disambiguating rare senses. Automatically extracted from Wiktionary, FEWS provides a large training set covering numerous senses and an evaluation set with few- and zero-shot examples, facilitating research in low-shot WSD scenarios in English.

WordNet (Miller, 1995) serves as a comprehensive lexical database grouping words into synsets representing distinct concepts. Each synset is interconnected through various semantic relations, offering a structured sense inventory integral to WSD tasks. It primarily focuses on English, but various projects have extended it to other languages.

BabelNet (Navigli and Ponzetto, 2012) extends WordNet by integrating it with Wikipedia and other resources, forming a multilingual semantic network. As of version 5.3 (December 2023), BabelNet covers 600 languages, containing almost 23 million synsets and around 1.7 billion word senses (Navigli et al., 2023). This expansive resource connects concepts across languages, supporting cross-lingual WSD and enriching the sense inventory beyond monolingual constraints.

2.3.2 WiC Datasets

The Word-in-Context (WiC) framework has been instrumental in evaluating context-sensitive word embeddings through binary classification tasks. Several notable datasets have been developed within this framework:

WiC (Pilehvar and Camacho-Collados, 2019b) is the pioneering English dataset that introduced the WiC framework. It consists of sentence pairs where a target word appears in both contexts, and the task is to determine whether the word carries the same meaning in both sentences. This dataset has set the standard for subsequent WiC-based evaluations.

XL-WiC (Raganato et al., 2020) extends the WiC framework to a multilingual setting, encompassing 12 languages: Bulgarian, Danish, German, Estonian, Farsi, French, Croatian, Italian, Japanese, Korean, Dutch, and Chinese. This expansion facilitates cross-lingual evaluation of semantic contextualization and enables research into zero-shot transfer capabilities of multilingual models.

MCL-WiC (Martelli et al., 2021) offers datasets in English, Arabic, French, Russian, and Chinese. These were constructed by annotating sentences from native corpora, including BabelNet (Navigli and Ponzetto, 2012), the United Nations Parallel Corpus (Ziemski et al., 2016), and Wikipedia. The dataset achieved inter-annotator agreements of 0.95 and 0.9 for English and Russian, respectively, indicating high annotation quality.

AM^{2}iCo (Liu et al., 2021) presents a multilingual dataset pairing English with 14 target languages. Compiled from Wikipedia dumps of each language, it selects words with at least two distinct pages, indicating ambiguity in both the target language and English. The dataset reports an overall human accuracy of 90.6% and an inter-annotator agreement of 88.4%.

WiC-TSV (Breit et al., 2021) introduces a multi-domain evaluation benchmark for WiC, independent of external sense inventories, but only in English. Covering various domains, WiC-TSV provides flexibility for evaluating diverse models and systems both within and across domains.

Despite these advancements, there remains a significant gap in resources for low-resource languages. Our dataset aims to address this deficiency by providing sense-annotated data in both WSD and WiC formats for underrepresented languages, thereby facilitating research in polysemy disambiguation and cross-lingual transfer across a broader spectrum of linguistic contexts.

3 Methods

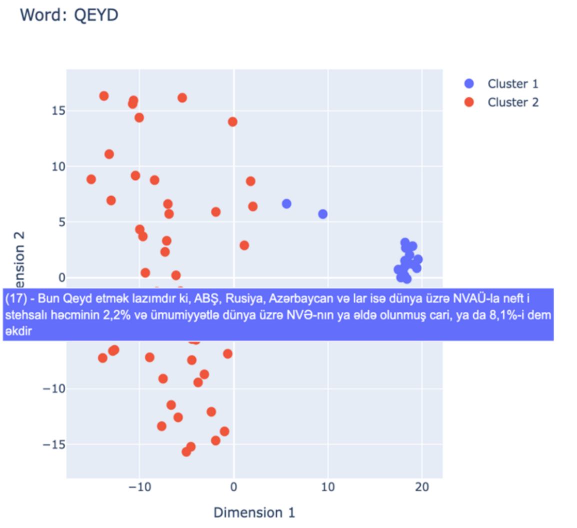
3.1 Dataset Curation

We follow the below method for the curation of sense-annotated datasets, adjusted for language-specific considerations. These are detailed in section §4.1, along with the resources used for the curation of the dataset in each language.

1. Identification of Polysemous Words Publicly available dictionaries (online and offline) were surveyed. By searching for words with more than a single dictionary entry, lists of hundreds of candidate polysemous words were compiled. Where available, lists of polysemous words were added.

2. Corpus Selection and Sentence Sampling Native corpora of sufficient size were chosen to ensure diverse contextual representation of target words. Candidate polysemous words were filtered based on corpus frequency, removing low-frequency terms, and manually reviewed for sense granularity. From these corpora, large samples of sentences (typically 100-1000 per word) were randomly extracted for further analysis.

3. Embedding-Based Analysis Word embeddings were generated for target words in the sampled sentences, and dimensionality reduction methods and clustering techniques were applied to these to create interactive 2D visualizations (see §3.2).



(17) - Bun Qeyd etmək lazımdır ki, ABŞ, Rusiya, Azərbaycan və lar isə dünya üzrə NVAÜ-la neft i stehsalı həcminin 2,2% və ümumiyyətlə dünya üzrə NVƏ-nın ya əldə olunmuş cari, ya da 8,1%-i dem əkdir

Figure 1: Example of interactive embedding-based sentence selection for the Azerbaijani word ‘qeyd’.

4. Manual Annotation of Sentences: In the 2D visualization, presented in Figure 1, annotators could hover over points representing sentences and click to assign them to different sense groups, for one word at a time. Sentences were selected based on their distribution in the embedding space or automatic clustering labels, with priority given to those that were more dispersed to ensure broad semantic coverage and enhance the representation of rare senses.

3.2 Semi-Automatic Annotation Tool

Our annotation process is semi-automatic, using vector representations for efficient sentence selection while ensuring manual verification.

To represent sentences in a structured way, we embed usages of the target word in all candidate sentences using pretrained transformer-based models such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), or language-specific models. These embeddings capture contextual semantics, making them suitable for sense-based clustering. We then apply K-Means or agglomerative clustering to group sentences into distinct senses, followed by dimensionality reduction techniques (e.g., UMAP, MDS) to visualize their distribution in 2D space (see Figure 1).

This visualization allowed annotators to interact with embeddings, exploring clusters and selecting diverse sentences that represent different word senses. This is essential for identifying sentences that correspond to rare word senses, as manually searching through randomly sampled sentences would be time-consuming and often ineffective, requiring the review of an extensive number of sentences to find relevant sentences.

3.3 Evaluating Annotation Efficiency

Annotating subordinate senses in polysemy is inherently time-consuming due to their rarity. Since these senses occur infrequently, manually identifying them requires scanning a large number of sentences before encountering a relevant instance.

The exact effort depends on the prior probability of the subordinate sense: the rarer it is, the more sentences need to be reviewed. To establish these priors, we randomly sampled 100 sentences for manual inspection to determine sense distributions. We then assessed how well model-based sentence selection captures each sense by comparing the proportion of automatically selected sentences correctly assigned to a sense against the baseline probability of encountering that sense in the corpus.

Our results demonstrate that computational methods significantly reduce this burden. We evaluate their effectiveness using adjusted **Lift**, a metric from Data Mining that measures improvement over random selection:

$$\text{Lift(sense)} = \frac{Precision(sense)}{Prior(sense)}$$

where *Precision(sense)* is the proportion of correctly classified sentences for the sense, and *Prior(sense)* their probability of occurrence in the dataset. Higher Lift values indicate a greater efficiency gain in selecting rare senses.

For example, in Kannada, identifying the word ಮತ in its subordinate 'religion' sense yielded a Lift of 900%, meaning that the likelihood of finding relevant sentences increased ninefold compared to random selection. Given a prior distribution of 96:4 favoring the dominant sense, manual selection would require reviewing 25 sentences on average to find one relevant case. With automatic selection achieving 36% precision, only three selections are needed—an 8× reduction in effort.

This efficiency boost translates directly into time and cost savings. If manual annotation takes 30 seconds per sentence, annotating 1,000 examples of a rare sense would traditionally require 8 hours of labor. With our automated method, this drops to about an hour, dramatically reducing annotation costs and making large-scale sense labeling more feasible. In Table 4, we present the Lift scores for the senses of two words in each of the four languages. Additional results, covering five words for each of these languages, are provided in Table B in the Appendix, covering all words selected for this evaluation.

Lang	**Word**	**Sense Definitions**		**Lift (%)**	
		1	2	1	2
KN	ಅಡಿ	Foot	Under	269	141
	ಮತ	Opinion	Religion	104	900
MR	रस	Juice	Interest	128	188
	उत्तर	Answer	North	121	125
PA	ਗੋਲੀ	Bullet	Pill	107	1364
	ਵਿਚਾਰ	Thought	Intention	235	884
UR	سونا	Gold	Sleep	161	232
	شکر	Thanks	Sugar	106	1414

Table 1: Measured improvement over random chance (Lift) in semi-automated sentence selection.

4 Sense-annotated Datasets

We introduce a sense-annotated corpus of sentences containing polysemous words covering ten low resource languages that span different language families and use different scripts: Azerbaijani (Turkic), Kannada and Telugu (Dravidian), Punjabi, Marathi and Urdu (Indo-Aryan), Polish (Slavic), Swahili (Afro-semitic), Vietnamese (Austroasiatic) and Korean (Koreanic). Statistics for each language are presented in Table 2.

4.1 Language Specific Treatment

For each language, the dataset was compiled and annotated by native speakers with the support of computational methods described above.

Azerbaijani: Polysemous words were selected from Azerbaycan Dilinin Omonimler Lugeti (Hesenov, 2007), and sentences containing selected target words were sampled from AzCorpus, the largest open-source NLP corpus for Azerbaijani (Kishiyev et al.). Three models were used to embed sentences: XLM-R, BERT-Turkish (DBMDZ, 2025), and XL-LEXEME (Cassotti et al., 2023).

Kannada: Polysemous words were selected from the online Kannada dictionary (Venkatasubbaiah et al., 1981). Kakwani et al. (2020a) was used as a corpus, which was preprocessed to remove extraneous characters, symbols, non-linguistic patterns, excessively long or single-word sentences, and duplicate entries. Initially, sentences for five words were annotated manually. Next, Claude 3.5 Sonnet (Anthropic, 2024) was used to pre-label

Language (ISO)	Words	Sentences	Senses	Avg. Senses/Word	Avg. Sentences/Sense
Azerbaijani (AZ)	60	4214	119	1.98 ± 0.13	35.55 ± 6.09
Kannada (KN)	59	4446	127	2.15 ± 0.45	35.01 ± 14.07
Korean (KO)	28	1013	58	2.07 ± 0.54	17.81 ± 4.73
Marathi (MR)	63	3766	125	1.98 ± 0.13	30.16 ± 2.72
Polish (PL)	66	2877	158	2.39 ± 0.68	18.28 ± 5.22
Punjabi (PA)	55	4969	127	2.31 ± 0.54	39.25 ± 1.89
Swahili (SW)	22	1376	46	2.09 ± 0.29	29.91 ± 4.39
Telugu (TE)	51	4534	100	1.96 ± 0.28	45.37 ± 7.83
Urdu (UR)	39	2674	90	2.31 ± 0.52	29.72 ± 1.06
Vietnamese (VI)	11	1021	29	2.64 ± 0.81	36.14 ± 19.20

Table 2: Statistics and ISO codes for the Multilingual WSD Sense-Annotated Dataset.

sentences after demonstrating reliable performance on the manually annotated data. The model, given Kannada and English meanings for each word, classified sentences containing the remaining target words. This streamlined human annotation, as annotators selected 30-40 sentences per sense from Claude's labels, rather than relying on clustering or embeddings for sentence selection. Finally, an independent reviewer verified all annotations.

Korean: The Korean Dictionary of National Institute of Korean Language (NIKL) (2025) was used to extract list of polysemous words. Two corpora were used for sampling sentences: the Korean Wikipedia Dataset (Lee, 2024) and KoWikiText (Kim, 2020). A Korean contextualized model (Ham et al., 2020) was used to embed sentences.

Marathi: The Marathi-English Dictionary from the Digital South Asia Library (DSAL) (Molesworth, 1857) was used to select polysemous words. For sampling sentences, three corpora were used: The Full Marathi Corpus (Joshi et al., 2022), and Marathi portions of two Indic corpora (Kakwani et al., 2020b; Kumar et al., 2023). MuRIL (Khanuja et al., 2021), mBERT, IndicBERT (Kakwani et al., 2020b), XLM-R, and XL-LEXEME were used for embedding sentences.

Polish: Polysemous words were identified by reviewing native texts, verified using the Polish Online Dictionary (Wydawnictwo Naukowe PWN, 2025), and selected if they had distinct senses. Three corpora covering distinct domains—national corpus, news, and literature—were used to sample sentences (Degórski and Przepiórkowski, 2012; Collection, 2018; Lebedev, 2023). XL-LEXEME and a Polish BERT (Kłeczek, 2020) were used for embedding sentences. Given Polish's high degree of inflection-where nouns, adjectives, and verbs vary by case, number, gender, and aspect across seven grammatical cases-all corpora were lemmatized to find sentences with target words in their base form for sentence selection and then restored to their original form for manual annotation.

Punjabi: Only text in Gurmukhi script was considered. Polysemous words were selected from previous work on WSD in Punjabi (Singh and Kumar, 2018, 2019, 2020; Singh and Singh, 2015) as well as from dictionaries (Joshi, 2009; Goswami, 2000; Brothers, 2006). Sentences were sampled from Metatext (Conneau et al., 2020), Samanantar (Ramesh et al.) and Sangraha (Khan et al.). MuRIL, IndicBERT, mBERT, XLM-R and XL-LEXEME were used to embed the sentences.

Swahili: The Swahili Dictionary (Chuo Kikuu cha Dar es Salaam, Taasisi ya Taaluma za Kiswahili, 2013) was used to identify polysemous words, while the Swahili Corpus by Masua and Masasi (2024) provided sentences. Multiple models were used for embedding (XLM-R, BERT and mBERT), but SwahBERT (Martin et al., 2022) outperformed them on the initial annotated dataset and was used to aid further annotation.

Telugu: Three corpora were used for selecting polysemous words, two Indic corpora (Kunchukuttan et al., 2020; Kakwani et al., 2020b) and the corresponding Wikipedia Dump (Wikimedia Foundation, 2024). The same Indic corpus (Kunchukuttan et al., 2020) was used for sentence selection, along with the Leipzig Telugu Corpus (Leipzig Corpora Collection, 2017). For embeddings, TeluguBERT (Joshi, 2022) and MuRIL were compared, with the former outperforming.

Urdu: Two word sense-annotated corpora (Saeed et al., 2019b,a), the Urdu Wiktextract (Ylonen, 2022), and a publicly available vocabulary book (Bruce, 2021) were used to select polysemous words. The Urdu Monolingual Corpus (UrMono) (Jawaid et al., 2014) was used to sample sentences. For embedding, mBERT, XLM-R, MuRIL, and XL-LEXEME were tested with the latter outperforming the rest. Given Urdu's complex inflectional morphology and honorific system, a list of up to six inflected forms was generated for each noun, considering variations in number, gender, and case to ensure a diverse sentence selection.

Vietnamese: Polysemous words were selected from the Tuttle Concise Vietnamese Dictionary (Giuong, 2014), while sentences containing target words were sampled from the English-Vietnamese Parallel Corpus (EVBCorpus) (Ngo et al., 2013). For embedding, XL-LEXEME, XLM-R, mBERT, as well as two Vietnamese-specific models, PhoBERT (Nguyen and Tuan Nguyen, 2020) and ELECTRA (Nguyen, 2025) were evaluated. As with other languages, PhoBERT emerged as the best model, highlighting the need for language-specific methods and resources.

4.2 WiC Pairing

For model training we convert the sense-annotated data in each language to the WiC format (see §2.2).

To guarantee that the train-dev-test splits contain well-representative samples of words and sentences, and ensure sentences appear only in a single split, we use the following steps to convert sense-annotated sentences to WiC sentence pairs:

1. Word Splitting 70% of the words are randomly allocated to the training set, while 15% each are allocated to validation and test sets.

2. Sentence Redistribution 30% of words from the training set are randomly selected to appear in all three splits (each sentence appearing only in one of the splits). For these words, 25% of their sentences are reallocated to the validation and test sets, ensuring: (1) Equal distribution between sets; (2) No sentence overlap across splits; and (3) The distribution of senses remains unchanged.

3. Pairing Sentences into WiC Pairs Within each split, each sentence is paired with up to 16 different sentences, ensuring a balanced mix of same-sense and different-sense pairs.

The amounts were selected to approximate a 70-15-15 dataset split. This approach ensures a representative, well-distributed, and balanced dataset for WiC training and testing, although it's important to note that different random seeds for sampling can result in different results, especially for smaller datasets. Descriptive statistics of the resulting WiC datasets can be found in Table 5 in the Appendix. All sets are approximately balanced, setting chance performance close to 50%.

5 Experiments

To assess the quality of the datasets we created, and to demonstrate the need for proper evaluation in low-resource languages, we tested transfer in three transfer conditions, full-shot, zero-shot and mixed. The **full-shot** condition is mainly a sanity-check, and serves to evaluate the quality of the training set, as it does not test for transfer. In **zero-shot**, a model is fine-tuned on English (combined training data taken from the MCL (Martelli et al., 2021) and XL (Raganato et al., 2020) datasets, totaling 13.4k sentence pairs) and evaluated on each of our ten languages, which it was not fine-tuned on. In the **mixed** condition, a model is first fine-tuned on English, and then on the target language training data, evaluating on the target language. This allows us to investigate whether leveraging large amounts of data in a high-resource language can enhance full-shot performance on low-resource corpora.

We use XLM-RoBERTa (Conneau et al., 2020) due to its strong multilingual capabilities. The model is pretrained on 100 languages, including all those in our novel datasets. It has proven highly effective in embedding both high- and low-resource languages and is widely studied in cross-lingual transfer research (Philippy et al., 2023), particularly in the context of polysemy disambiguation (Raganato et al., 2020; Dairkee and Dubossarsky, 2024; Cassotti et al., 2023).

For model fine-tuning, we follow Cassotti et al. (2023) and use a bi-encoder architecture that independently processes the two sentences containing the polysemous target word using a Siamese network to generate two distinct vector representations (embeddings). The model outputs the cosine distance between the output embeddings of the two inputs, and, to collapse this to a binary label, a threshold is applied to decide if the words are classified as having the same sense. The model is trained to adapt embeddings and increase this distance when the target word has different meanings and decrease it when the meanings are the same in the two sentences by minimising contrastive loss. After training, we set the threshold for each model

Test Lang Condition	AZ	KN	KO	MR	PL	PA	SW	TE	UR	VI	Avg.
Full-shot	65.9	65.9	56.4	83.2	72.3	65.9	59.5	63.8	68.8	57.2	65.9
Zero-shot	66.3	**72.3**	64.2	82.2	**79.1**	70.5	68.6	62.4	**74.0**	**70.6**	71.0
Mixed	**71.9**	71.0	**66.5**	**88.1**	65.4	**81.6**	**76.9**	**65.4**	64.8	68.4	**72.0**

Table 3: Accuracies of XLM-R models evaluated on the test sets of our WiC datasets. Full-shot refers to models trained exclusively on the target language's training data. Zero-shot results correspond to XLM-R trained only on English WiC data. Mixed models are first trained on English, then fine-tuned on the target language.

by maximising accuracy on the corresponding validation set. During training, as well as inference, special tokens, `<t>` and `</t>`, are placed around the target word in each sentence to signal what word the model should focus on.

6 Results

Our semi-automatic annotation method works The transfer results (Table 3) demonstrate that we were able to produce high-quality datasets in ten low-resource languages. The low performance in Korean, Swahili, and Vietnamese is only observed in the full-shot condition. These are most likely due to their smaller training size rather than quality issues; otherwise, low performance would have been observed also in the zero-shot condition.

Evaluating on all target languages is essential Transfer effects are not uniform, as seen in the zero-shot performance that varies from 62.4% in Telugu to 82.2% in Marathi. Interestingly, zero-shot outperforms full-shot in 8 out of 10 languages, and gets comparable accuracy in the remaining 2, likely due to the small training data size of full-shot models and strong transfer from English. These results emphasize the unpredictability of transfer from one side, but also stress the need for a comprehensive multilingual benchmark to accurately assess cross-lingual transfer and ensure models perform reliably across diverse languages. With our efficient semi-automatic annotation method, curating such datasets is also much cheaper in annotation efforts.

Mixed training improves transfer For most languages, mixed-training improves upon either full-shot or zero-shot conditions. This hybrid strategy leverages large-scale training data in English with language-specific details from the target language for effective polysemy resolution. This further highlights the importance of datasets in low-resource languages, where even small amounts of labeled data can lead to marked improvements.

7 Discussion

In this work we present sense-annotated datasets across a diverse range of language families, providing valuable resources for linguistic and computational studies. Punjabi, Marathi, and Urdu belong to the Indo-Aryan branch, enabling research on linguistic relatedness alongside the Hindi WiC dataset (Dairkee and Dubossarsky, 2024). Telugu and Kannada represent the Dravidian family, while Azerbaijani, Swahili, Vietnamese, Polish, and Korean extend coverage to additional linguistic groups. The dataset includes Arabic-based (Punjabi, Urdu), Devanagari (Marathi), Latin-based (Azerbaijani, Polish, Swahili, Vietnamese), Hangul (Korean), and Brahmic scripts (Kannada, Telugu), facilitating research on script variation and its impact on NLP.

By encompassing a broad linguistic spectrum, our dataset supports studies on linguistic relatedness, historical evolution, and polysemy disambiguation in low-resource settings. It serves as a foundation for evaluating and improving multilingual and cross-lingual transfer, particularly in tasks requiring deep semantic understanding.

Our experiments highlight the importance of language-specific resources. The unexpected finding that zero-shot XLM-R trained only on English outperformed full-shot models trained on the target language challenges assumptions about cross-lingual transfer stability, emphasizing the need for dedicated evaluation datasets.

Manual annotation is essential yet labor-intensive, particularly for low-resource languages. We introduce an automated method to identify sentences across all word senses, even when certain senses are sparsely represented. Our quantitative results demonstrate the effectiveness of this approach in enhancing annotation efficiency and supporting sense-annotated dataset development. To encourage further research, we release our code on GitHub: github.com/roksanagow/projecting_sentences.

8 Limitations

The dataset remains relatively small, which may limit the generalizability of findings, particularly for full-shot experiments, where additional training data would likely improve performance. Additionally, data imbalance across languages makes direct comparisons challenging without subsampling, which in turn reduces overall performance. Even within a single language, the number of senses and sentences per word varies, further complicating evaluation. Moreover, each language was sourced from different corpora, leading to potential inconsistencies in text style, domain coverage, and annotation quality.

The evaluation setup also has certain constraints. Train-dev-test splits were generated randomly (according to the algorithm specified in §4.2), and the prevalence of sentences corresponding to different words across splits could impact the results. Furthermore, zero-shot evaluation was conducted only from English, leaving open questions about transfer from other high-resource languages and cross-lingual settings beyond English-centric transfer.

9 Acknowledgments

This work was partially funded by the research program Change is Key!, supported by Riksbankens Jubileumsfond (reference number M21-0021). The authors would like to thank Bao Linh Hoang for contributing additional expert annotations.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Google AI. 2018. Multilingual bert (mbert). Accessed: August 2024.

Anthropic. 2024. Claude 3.5 sonnet. Available at https://www.anthropic.com/news/claude-3-5-sonnet.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *International Joint Conference on Artificial Intelligence*, pages 4330–4338. International Joint Conference on Artificial Intelligence, Inc.

Terra Blevins, Mandar Joshi, and Luke Zettlemoyer. 2021. FEWS: Large-scale, low-shot word sense disambiguation with the dictionary. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2554–2560. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Singh Brothers. 2006. *Punjabi-English Dictionary*. Singh Brothers, Amritsar. Accessed: August 2024.

Gregory Maxwell Bruce. 2021. *Urdu Vocabulary: A Workbook for Intermediate and Advanced Students*. Edinburgh University Press, Edinburgh. Accessed: August 2024.

Agostina Calabrese, Michele Bevilacqua, Roberto Navigli, et al. 2020. Fatality killed the cat or: Babelpic, a multimodal dataset for non-concrete concepts. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4680–4686. Association for Computational Linguistics.

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.

Nadezhda Chirkova and Vassilina Nikoulina. 2024. Zero-shot cross-lingual transfer in instruction tuning of large language models. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 695–708, Tokyo, Japan. Association for Computational Linguistics.

Chuo Kikuu cha Dar es Salaam, Taasisi ya Taaluma za Kiswahili. 2013. *Kamusi ya Kiswahili Sanifu*. Oxford University Press, East Africa Limited, Nairobi, Kenya. Accessed: August 2024.

Leipzig Corpora Collection. 2018. pol_news_2018: Polish news corpus. `https://corpora.wortschatz-leipzig.de/ic?corpusId=pol_news_2018`. Accessed: August 2024.

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.

Farheen Dairkee and Haim Dubossarsky. 2024. Strengthening the wic: New polysemy dataset in hindi and lack of cross lingual transfer. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15341–15349.

DBMDZ. 2025. BERT-Base Turkish 128K Uncased. Available at Hugging Face, Accessed: August 2024.

Łukasz Degórski and Adam Przepiórkowski. 2012. Recznie znakowany milionowy podkorpus nkjp. *Przepiórkowski et al.[17]*, pages 51–58.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*, arXiv:1810.04805.

Błażej Dolicki and Gerasimos Spanakis. 2021. Analysing the impact of linguistic features on cross-lingual transfer. *arXiv preprint arXiv:2105.05975*.

Philip Edmonds and Scott Cotton. 2001. Senseval-2: Overview. In *Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5. Association for Computational Linguistics.

Phan Van Giuong. 2014. *Tuttle Concise Vietnamese Dictionary: Vietnamese-English, English-Vietnamese*. Tuttle Publishing, North Clarendon, VT. Accessed: August 2024.

K. K. Goswami. 2000. *Punjabi-English/English-Punjabi Dictionary*. Hippocrene Books, New York. Accessed: August 2024.

J. Ham, Y. J. Choe, K. Park, I. Choi, and H. Soh. 2020. Ko-sroberta multitask model. Available at Hugging Face, Accessed: August 2024.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2021. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*.

Hesret Hesenov. 2007. Azerbaycan dilinin omonimler lugeti. *Serq Qerb nesriyyati. Baki*.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60. Association for Computational Linguistics.

Bushra Jawaid, Amir Kamran, and Ondřej Bojar. 2014. A tagged corpus and a tagger for urdu. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2938–2943, Reykjavik, Iceland. European Language Resources Association (ELRA). Accessed: August 2024.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*. `https://arxiv.org/abs/2211.11418`.

Raviraj Joshi et al. 2022. L3Cube-MahaNLP: Marathi natural language processing datasets, models, and library. Accessed: August 2024.

S. S. Joshi. 2009. *Punjabi-English Dictionary: Panjabi Yuniwarasiti Panjabi-Angarezi Kosha*. Punjabi University, Patiala. Accessed: August 2024.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020a. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020b. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961. Accessed: August 2024.

Mohammed Safi Ur Rahman Khan, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Suriyaprasaad G, Varun Balan G, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, and Mitesh M. Khapra. IndicLLMSuite: A Blueprint for Creating Pre-training and Fine-Tuning Datasets for Indian Languages.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages. *Preprint*, arXiv:2103.10730.

Hyunjoong Kim. 2020. Kowikitext: A wikitext format korean corpus. Accessed: August 2024.

Huseyn Kishiyev, Jafar Isbarov, Kanan Suleymanli, Khazar Heydarli, Leyla Eminova, and Nijat Zeynalov. azcorpus: The largest open-source nlp corpus for azerbaijani (1.9m documents, 18m sentences). Accessed: August 2024.

Anoop Kumar et al. 2023. Sangraha: A high-quality, multilingual dataset for indic language pretraining. Accessed: August 2024.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. AI4Bharat-IndicNLP Corpus: Monolingual Corpora and Word Embeddings for Indic Languages. *arXiv preprint arXiv:2005.00085*. Accessed: August 2024.

Darek Kłeczek. 2020. Polbert: Polish bert language model. Accessed: August 2024.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Dmitrii Lebedev. 2023. Polish classic literature text corpus. Accessed: August 2024.

Chang W. Lee. 2024. Wikipedia korean dataset (2024-05-01). Accessed: August 2024.

Leipzig Corpora Collection. 2017. Telugu community corpus (2017). Accessed: August 2024.

Qianchu Liu, Edoardo Maria Ponti, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2021. AM2iCo: Evaluating word meaning in context across low-resource languages with adversarial examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7151–7162, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics*, 47(2):387–443.

Federico Martelli, Najla Kalach, Gabriele Tola, Roberto Navigli, et al. 2021. Semeval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (mcl-wic). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 24–36.

Gati Martin, Medard Edmund Mswahili, Young-Seob Jeong, and Jiyoung Woo. 2022. SwahBERT: Language model of Swahili. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 303–313, Seattle, United States. Association for Computational Linguistics.

Bernard Masua and Noel Masasi. 2024. In the heart of swahili: An exploration of data collection methods and corpus curation for natural language processing. *Data in Brief*, 55:110751.

George A. Miller. 1995. WordNet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross Bunker. 1993. A semantic concordance. In *Proceedings of the Workshop on Human Language Technology*, pages 303–308. Association for Computational Linguistics.

James Thomas Molesworth. 1857. *A Dictionary, Marathi and English*, 2 edition. Printed for government at the Bombay Education Society's Press, Bombay. Accessed: August 2024.

National Institute of Korean Language (NIKL). 2025. Nikl korean-english dictionary. Dataset available on Hugging Face. Accessed: August 2024.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. Semeval-2013 task 12: Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, Atlanta, Georgia, USA. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Roberto Navigli et al. 2023. BabelNet 2023. `https://babelnet.org/publications`.

Quoc Hung Ngo, Werner Winiwarter, and Bartholomäus Wloka. 2013. EVBCorpus - a multi-layer english-vietnamese bilingual corpus for studying tasks in comparative linguistics. In *Proceedings of the 11th Workshop on Asian Language Resources (ALR-11) at IJCNLP 2013*, pages 1–9, Nagoya, Japan. Asian Federation of Natural Language Processing. Accessed: August 2024.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online. Association for Computational Linguistics.

Nha Van Nguyen. 2025. Nlphust/ner-vietnamese-electra-base. Accessed: August 2024.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019a. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019b. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 282–293, Brussels, Belgium. Association for Computational Linguistics.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. Semeval-2007 task-17: English lexical sample, srl and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic. Association for Computational Linguistics.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. Xl-wic: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. 10:145–162.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Christoph Rzymski, Tiago Tresoldi, Simon J Greenhill, Mei-Shin Wu, Nathanael E Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A Bodt, Abbie Hantgan, Gereon A Kaiping, et al. 2020. The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific data*, 7(1):13.

Ali Saeed, Rao Muhammad Adeel Nawab, Mark Stevenson, and Paul Rayson. 2019a. A sense annotated corpus for all-words urdu word sense disambiguation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 18(4):1–14. Accessed: August 2024.

Ali Saeed, Rao Muhammad Adeel Nawab, Mark Stevenson, and Paul Rayson. 2019b. A word sense disambiguation corpus for urdu. *Language Resources and Evaluation*, 53(3):397–418. Accessed: August 2024.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2304–2317, Bangkok, Thailand. Association for Computational Linguistics.

Jagbir Singh and Iqbal Singh. 2015. Word sense disambiguation: Enhanced lesk approach in punjabi language. *International Journal of Computer Applications*, 129(6):23–27. Accessed: August 2024.

Varinder Pal Singh and Parteek Kumar. 2018. Naive bayes classifier for word sense disambiguation of punjabi language. *Malaysian Journal of Computer Science*, 31(3):188–199. Accessed: August 2024.

Varinder Pal Singh and Parteek Kumar. 2019. Sense disambiguation for punjabi language using supervised machine learning techniques. *Sādhanā*, 44(11):226. Accessed: August 2024.

Varinder Pal Singh and Parteek Kumar. 2020. Word sense disambiguation for punjabi language using deep learning techniques. *Neural Computing and Applications*, 32(8):2963–2973. Accessed: August 2024.

Anirudh Srinivasan, Sunayana Sitaram, Tanuja Ganu, Sandipan Dandapat, Kalika Bali, and Monojit Choudhury. 2021. Predicting the performance of multilingual nlp models. *arXiv preprint arXiv:2110.08875*.

Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. 2021. BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies? 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 3609–3624, Online. Association for Computational Linguistics.

G. Venkatasubbaiah, L. S. Sheshagiri Rao, and H. K. Ramachandra. 1981. *Kannada-Kannada-English Dictionary*. Ibh Prakashana, Bangalore, India. Accessed: August 2024.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Wikimedia Foundation. 2024. Telugu wikipedia dump. Accessed: August 2024.

Wikimedia Foundation. 2025. List of wikipedias. Accessed: August 2024.

Wydawnictwo Naukowe PWN. 2025. Słownik języka polskiego pwn. Accessed: August 2024.

Tatu Ylonen. 2022. Wiktextract: Wiktionary as machine-readable structured data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1317–1325, Marseille, France. European Language Resources Association. Accessed: August 2024.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

A Example visualization of annotated sentences

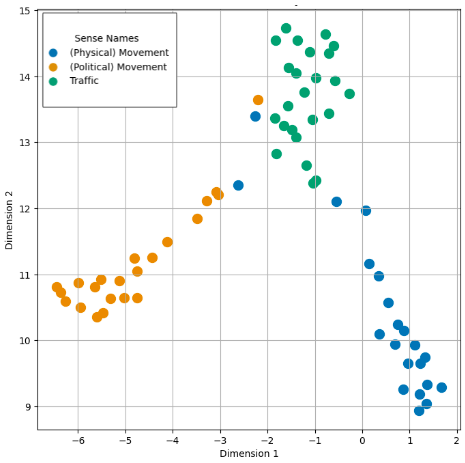


Figure 2: Word embeddings of the Polish word 'ruch' in sense-annotated sentences, visualized in 2D with UMAP. Interestingly, the resulting shape resembles a walking figure.

B Evaluating Annotation Efficiency

Lang	**Word**	**Sense Definitions**			**Lift (%)**		
		1	2	3	1	2	3
KN	ಅಡಿ	Foot	Under	-	269	141	-
	ಮತ	Opinion	Religion	-	104	900	-
	ಗುಂಡಿ	Pit/Hole	Bullet	-	269	141	-
	ಮಂಡಿ	Market	Knee	-	167	223	-
	ಮಾತು	Word	Conversation	-	289	128	-
MR	रस	Juice	Interest	-	228	288	-
	उत्तर	Answer	North	-	221	225	-
	मान	Respect	Approval	-	218	235	-
	खोली	Room	Depth	-	147	124	-
	हार	Necklace	Defeat	-	106	149	-
PA	ਗੋਲੀ	Bullet	Pill	-	107	1364	-
	ਵਿਚਾਰ	Thought	Intention	-	235	884	-
	ਉੱਤਰ	North	Response	Descend	210	438	156
	ਖਾਨ	Khan (name)	Mine	-	128	211	-
	ਹਾਰ	Defeat	Necklace	-	129	∞ (prior = 0)	-
UR	سونا	Gold	Sleep	-	161	232	-
	شکر	Thanks	Sugar	-	106	1414	-
	زبان	Language	Tongue	-	119	358	-
	کانٹا	Thorn	Fork	-	108	808	-
	اتفاق	Opportunity	Agreement	Coincidence	685	155	364

Table 4: Measured improvement over random chance (Lift) in semi-automated sentence selection over all evaluated words.

C WiC sentence pairing

Language	AZ	KN	KO	MR	PL	PA	SW	TE	UR	VI
Sent Pairs (Train)	20,409	20,298	5,703	19,368	13,516	26,237	7,312	23,115	14,018	5,153
Sent Pairs (Dev)	5,649	5,627	1,018	5,175	3,562	7,025	2,165	5,861	3,450	751
Sent Pairs (Test)	5,434	4,809	656	4,194	3,103	5,749	1,100	5,500	3,210	1,397
Words (Train)	42	42	20	45	47	39	16	36	28	8
Words (Dev)	22	22	11	24	25	21	9	19	15	5
Words (Test)	22	21	9	22	24	19	7	18	14	4
Words in All Splits	13	13	6	14	15	12	5	11	9	3

Table 5: Amounts of sentence pairs and unique polysemous target words in the train-dev-test splits of our constructed WiC datasets.