

# Generating Concept Lexicalizations via Dictionary-Based Cross-Lingual Sense Projection

David Basil\*, Chirooth Girigowda, Bradley Hauer, Sahir Momin, Ning Shi, Grzegorz Kondrak\*  
Alberta Machine Intelligence Institute (Amii)  
Department of Computing Science  
University of Alberta, Edmonton, Canada

## Abstract

We study the task of automatically expanding WordNet-style lexical resources to new languages through sense generation. We generate senses by associating target-language lemmas with existing lexical concepts via semantic projection. Given a sense-tagged English corpus and its translation, our method projects the annotated synsets onto aligned target-language tokens and assigns the corresponding lemmas to those synsets. To generate alignments and ensure their quality, we augment a pretrained base aligner with a bilingual dictionary, which is also used to filter incorrect sense projections. We evaluate the method on multiple languages, comparing it to prior methods, as well as dictionary-based and large language model baselines. Results show that the proposed project-and-filter strategy improves precision while remaining interpretable and resource-efficient. We release our code, documentation, and generated sense inventories at <https://github.com/UAlberta-NLP/ExpandNet>.

**Keywords:** Lexicon, Lexical Semantics, Multilinguality, WSD, Wordnets.

## 1. Introduction

A wordnet is a semantic resource that organizes words into synonym sets, or *synsets*, providing a structured way to represent word meaning. Each synset represents a distinct lexical concept and contains the lemmas that express it. Wordnets are widely used throughout natural language processing. Despite their importance, automated wordnet construction remains challenging, particularly in the multilingual setting, where words from different languages must be organized into a shared semantic ontology. A central component of this process is sense generation [1, 2], which is equivalent to adding lemmas to existing synsets.

We propose a sentence-level method for generating senses in a target language through dictionary-guided projection. Building on the semantic projection framework, we derive these new senses from translations of a sense-annotated source text. To ensure high-quality word alignment, we augment a pretrained aligner with a bilingual dictionary. We also use this bilingual dictionary to filter semantic projections, reducing erroneous synset assignments. To our knowledge, this represents the first application of dictionary-based filtering to sense generation. Our projection algorithm functions strictly at the sentence level, successfully eliminating the reliance on corpus-level statistics found in prior work. Rather than operating completely independently, we envision our system as a high-precision recommendation tool which provides candidate senses reliable enough for efficient verification by a native speaker.

Our project-and-filter procedure rests on three assumptions about word senses and translation: (1) Each content word in context expresses a single lexical concept. (2) Word senses are represented in a multilingual wordnet as synset-lemma pairs, each corresponding to a single lexical concept. (3) When the alignment of words in a sentence and its translation is supported by lexical evidence (e.g., a bilingual dictionary or a clear cognate relationship), the alignment represents a literal translation, and the aligned words express the same concept.

An extended version of this paper is available at <https://arxiv.org/abs/2604.14397>.

\*{dbasil1, gkondrak}@ualberta.ca

We expect our method to be particularly useful in settings where comprehensive wordnets or sense inventories are not yet available, including many low-resource languages. While the semantic projection framework inherently assumes access to a bitext, we generate this automatically by translating a sense-annotated source corpus into the target language. Our pipeline relies on machine translation, a bilingual dictionary (for both alignment and filtering), and an optional part-of-speech (POS) tagger. Critically, these resources are generally more accessible than sense inventories in many languages.

Experiments on four diverse language pairs demonstrate that our approach improves precision compared to prior projection-based methods. Dictionary filtering drives this precision by reducing spurious sense assignments, while our dictionary-augmented aligner improves coverage by introducing links that might otherwise be missed. Overall, the method achieves competitive performance while remaining highly interpretable. Furthermore, our error analysis suggests that most remaining errors arise from noisy resources or upstream mistakes in translation or alignment, rather than from limitations of our methodology itself.

## 2. Method

Our method, **ExpandNet**, builds upon the semantic projection framework, in which sense tags from a source-language corpus are transferred to a target language through word alignment. Because each sense tag corresponds to a lexical concept, projection associates target-language lemmas with existing concepts, thereby generating new senses in the target language. This framework assumes a word-aligned parallel corpus with sense annotations on the source side. Formally, if a source word  $x$  tagged with concept  $s$  is aligned with a target word  $y$ , then the lemma of  $y$  can also be associated with concept  $s$ . We say that the semantic tag  $s$  is *projected* across the alignment link from  $x$  to  $y$ , which *generates* the sense  $(y, s)$ .

Our method incorporates a filtering step that discards projections lacking dictionary support. To validate each aligned pair, the filtering step uses a bilingual dictionary, consisting of translation pairs of words or phrases in the source and target languages that serve as valid translations of one another. If a sense-tagged word is aligned with a target word, we project the sense if and only if the translation pair is found in the dictionary. To facilitate the generation of senses involving rare words and proper nouns, we always project a sense when the aligned words share the same orthographic form. We optionally apply a filter which discards projections when the source and target words do not share the same POS category (noun, verb, adjective, or adverb). Table C1 details the mapping of multiple tagsets to these four broad categories. Words falling outside these categories are excluded from projection.

Our method employs an aligner, **DBAlign**, which augments a pre-existing base aligner with evidence from a bilingual dictionary. This integration improves the robustness of the base aligner’s output by prioritizing dictionary-supported alignments and explicitly handling multi-word expressions. At a high level, **DBAlign** chooses a set of alignment links through a three-pass procedure. First, it accepts the intersection of links proposed by both the dictionary and the base aligner. We treat these alignments as high-confidence, as they are supported by both sources. Next, it incorporates any remaining dictionary-suggested links that do not conflict with previously accepted alignments. Finally, it adds remaining base aligner links that do not conflict with the previously accepted set. This ordering prioritizes high-confidence alignments before increasing coverage while avoiding conflicts. Throughout this process, two links are considered to conflict if they share a source or target token. If a given token has multiple candidate links that satisfy the condition of the current pass, we resolve the tie using the *diagonal heuristic*, adapted from the `fast_align` system [3]. The heuristic selects the link that minimizes the difference between the relative positions of the source and target tokens within their respective sentences.

### 3. Experiments

In this section, we describe our evaluation framework and the systems used for comparison, present results across four language pairs, and conclude with error analysis and an ablation study. The experimental setup and resources are detailed in Appendix A.

#### 3.1. Evaluation

When evaluating generated senses, the objective is to determine whether a proposed lemma-synset pairing is valid, that is, whether the lemma can lexicalize the corresponding concept. Evaluating generated senses remains a challenging open problem [4]. As an actively developed, well-established multilingual wordnet, BabelNet [5] serves as the gold standard in our automatic evaluation. Accordingly, a pairing of a word with a BabelNet synset by a sense generation system is deemed correct if and only if the synset contains the word. For all experiments, we use the latest release, BabelNet 5.3.

We report the performance of sense generation methods using two metrics. For *Sense Precision*, we calculate the proportion of the proposed senses that are found in BabelNet. For *Synset Coverage*, we calculate the proportion of synsets for which the method successfully proposes at least one correct sense. For both metrics, we restrict our evaluation to BabelNet synsets which contain at least one target-language lemma. This restriction aims to account for the incomplete coverage in BabelNet, since any proposed sense mapped to an empty synset would necessarily be deemed incorrect. We argue that this formulation is appropriate because the absence of lemmas in a synset for a given language is more likely to reflect a missing entry in BabelNet than a genuine lexical gap in the language itself. This distinction is crucial for cross-lingual comparison, as BabelNet coverage varies substantially across languages: Within our dataset, 93.9% of synsets contain at least one Spanish lemma, compared to 83.7% for French, 75.0% for Chinese, and only 17.5% for Urdu.

We consider precision to be the most important measure of performance. In sense generation, precision is vital because missed projections result in omissions that can be mitigated by processing larger corpora, making false positives the only source of incorrect entries in the resulting output. We do not report sense *recall* because there is no principled way to define a comprehensive “ground set” of all correct target-language senses that could theoretically be projected given a corpus. Such a ground set cannot be derived from the sense-tagged source corpus, as a single source-language sense may be translated in multiple ways, including non-literally. Instead, we use *Synset Coverage* as a proxy for recall. Its theoretical upper bound is reached when at least one correct sense is projected for every non-empty synset observed in the source corpus. While no projection-based method is expected to achieve full coverage, this measure, unlike the raw count of correctly generated senses, enables meaningful comparison across different datasets and target languages.

#### 3.2. Comparison Systems

We evaluate *ExpandNet* against two baselines: a dictionary-based approach and a direct large language model (LLM) generation strategy. The dictionary baseline naively assumes that, for a given sense in the source-side corpus, any target-language translation of the source word can express the corresponding meaning, effectively ignoring all contextual information. For the second baseline, we prompt an LLM to generate a target-language lexicalization for each concept based on its definition. This lexicalization is then paired with the concept to create a single sense. Specifically, we use `gpt-5-chat-latest` with the prompt template shown in Figure B1. By design, this baseline produces exactly one sense per synset.

The Universal WordNet (UWN) [6] is the earliest attempt to automatically expand the Princeton WordNet to cover a large set of additional languages. UWN does not employ a

Method	French (dev)			Spanish			Chinese			Urdu		
	Prec.	Cov.	HM	Prec.	Cov.	HM	Prec.	Cov.	HM	Prec.	Cov.	HM
UWN	46.1	56.9	50.9	71.9	51.4	60.0	45.0	5.8	10.3	<b>59.9</b>	13.1	21.5
Dict	5.2	<b>93.7</b>	9.9	8.5	<b>88.2</b>	15.5	7.8	<b>75.1</b>	14.1	9.4	44.0	21.5
OC14	71.1	8.2	14.6	72.0	9.9	17.4	34.3	2.8	5.2	44.0	0.5	1.0
M23	52.3	67.2	59.0	55.3	67.8	60.9	34.7	48.4	40.6	6.3	24.4	10.0
ChatGPT	44.3	44.3	44.3	55.3	55.3	55.3	41.8	41.8	41.8	50.5	<b>50.5</b>	<b>50.5</b>
ExpandNet	<b>71.5</b>	69.7	<b>70.6</b>	<b>82.8</b>	66.6	<b>73.8</b>	<b>64.8</b>	44.4	<b>52.7</b>	56.1	28.6	37.9

Table 1. Results (in%) on the SC20 corpus, showing the Sense Precision, Synset Coverage, and Harmonic Mean (HM) of the two.

projection framework; instead, it relies on a variety of resources, including a small amount of manually labeled training data, to train a supervised classifier capable of recognizing valid senses. We obtain UWN as provided by the authors in TSV format. From this resource, we extract all lemma-synset pairs associated with a “rel:means” relation where the lemma belongs to the target language. We then map UWN’s Princeton WordNet synset offsets to BabelNet synset IDs using the Python [BabelNet API](#). For every synset, we treat the corresponding target-language senses as the output.

Next, we re-implement the method from Oliver and Climent [1], which we refer to as OC14. This approach uses two parameters,  $i$  and  $f$ , to filter the generated senses. Parameter  $i$  specifies the minimum ratio between the highest and second-highest co-occurrence frequencies for any given synset. In this context, a target word  $y$  is said to “co-occur” with a synset  $s$  if  $y$  appears in a target sentence that is aligned with a source sentence containing a word tagged with  $s$ . Parameter  $f$  dictates the maximum allowable ratio of the synset’s frequency to the word’s frequency. A candidate sense is projected only if its respective ratios satisfy both thresholds. Following the original paper, we set  $i = 2.5$  and  $f = 5$ .

We also re-implement the method of Martelli, Procopio, Barba, and Navigli [2], which we refer to as M23. The method uses a single parameter,  $\beta$ , to threshold the total frequency of the target words retrieved for each source sense. The authors report results for  $\beta = 0.7$ ,  $\beta = 0.9$ , and  $\beta = 1.0$ . We use  $\beta = 0.7$  in our experiments, as this setting achieved the best performance among the three.

To provide a consistent comparison, we strive for uniform settings across all evaluated methods. In particular, our re-implementations of existing methods use the same translations as employed by our ExpandNet method, as well as SimAlign, the base aligner of DBAlign, where applicable. Since both OC14 and M23 make use of corpus-level statistics when deciding whether to project a sense annotation, applying them to the same corpus as ExpandNet ensures a fair comparison while demonstrating ExpandNet’s effectiveness even on limited data.

### 3.3. Results

Table 1 summarizes the results of our experiments. While the dictionary baseline achieves high coverage for most languages, its low precision makes it impractical for many use cases. This highlights the importance of context-aware translation and reinforces why bitexts are typically used for this task. The LLM baseline achieves fairly consistent performance across languages. Because it was prompted to provide only one lexicalization per synset, its coverage could likely be increased by allowing multiple senses. However, there is no reason to expect this would improve its precision, which is below that of ExpandNet in all languages. These results suggest that sense generation remains challenging for LLMs. Nevertheless, an LLM-based approach could serve as a useful back-off for senses that are not found in the

projected corpus, such as rare senses or those corresponding to lexical gaps in the source language. By definition, if a concept lacks a lexicalization in the source language, it cannot be projected to the target language. This is an inherent weakness of the projection framework, which direct LLM prompting can bypass.

Among methods from prior work, UWN achieves competitive coverage for French and Spanish, but its precision varies across languages. Coverage in Urdu and Chinese is poor, although UWN is more precise than **ExpandNet** in Urdu—the only case in which any system surpasses **ExpandNet** in precision. OC14 attains relatively high precision on French and Spanish, but exhibits the lowest coverage across all languages, making it ill-suited for populating a wordnet. The most recent comparison method, M23, achieves substantially better coverage, but at the cost of lower precision overall.

**ExpandNet** achieves strong performance across languages and metrics, often surpassing baselines and methods from prior work, particularly in precision. It generates 8,761 correct senses in Spanish, 7,365 in French, 4,768 in Chinese, and 630 in Urdu. These results also provide evidence for the effectiveness of **ExpandNet**'s sentence-level approach, which operates independently of corpus-level statistics, making it robust to variations in corpus size.

### 3.4. Error Analysis

While **ExpandNet** performs well relative to prior work, absolute scores remain moderate, particularly for Chinese and Urdu. This may actually reflect constraints in the evaluation setup, as a projection is counted as correct only if the projected sense appears in BabelNet, whose coverage varies substantially across languages. While we restrict evaluation to synsets that are not empty, even these may lack many valid lemmas, especially in languages that are typologically distant from English or are less densely represented in multilingual lexical resources. Consequently, automatic evaluation may incorrectly label valid projections as errors. In such cases, the system may in fact be proposing legitimate lexicalizations that are absent from the resource, suggesting that **ExpandNet** can serve not only as a tool for extending existing multilingual lexicons, but also as an evaluation target.

To investigate this issue more carefully, we conducted a targeted manual analysis. We focused on Chinese and randomly sampled 20 projected senses that were marked as incorrect due to their absence from BabelNet. Each case was reviewed by one of the authors of this article, a native Chinese speaker, who judged whether the projected lemma appropriately expressed the intended concept in context. Our analysis suggests that many apparent errors produced by **ExpandNet** are, in fact, valid projections that fall outside the coverage of BabelNet. Of the 20 sampled Chinese senses marked incorrect, 14 were judged to be completely valid. The remaining 6 cases primarily reflect noise in upstream tools and resources: three errors arose from inaccurate source-side sense annotations, two errors resulted from dictionary issues which allowed incorrect projections to pass the dictionary filter, and the final analyzed case involved a translation error.

The native speaker also analyzed cases involving sense-tagged tokens that were not aligned and thus not projected. Such unaligned tokens represent errors that decrease coverage. Ten unaligned lemmas were identified within the first 200 lemmas of the dataset. Of these, five were found to be cases of non-literal translations, where the sentence translation, while correct, does not have a direct equivalent for the source word in question.

In summary, many errors identified by automatic evaluation are attributable to gaps in BabelNet rather than incorrect projections. The remaining errors are primarily due to mistakes in annotation, alignment, or translation resources, rather than to flaws in the method itself. These findings are consistent with observations across other languages.

### 3.5. Ablation

We conducted an ablation study intended to quantify the impact of the principal innovations of **ExpandNet**, with Spanish as the target language. The results, summarized in Table D1, show that (1) the dictionary filter improves precision from 49.1% to 83.9%; (2) **DBAlign** increases coverage compared to **SimAlign**, from 59.5% to 68.3%; and (3) the best harmonic mean of precision and coverage is achieved with **DBAlign**, the dictionary filter, and no POS filter. Overall, these results indicate that the POS filter may not be essential, at least for Spanish, but we recommend using it when precision is a priority.

## 4. Conclusion

**ExpandNet** establishes a competitive standard for sense generation, delivering strong results across four language pairs and demonstrating scalability to small corpora. Our method achieves higher precision than previously published systems, without relying on corpus-level statistics to filter spurious alignments. Instead, it uses an innovative dictionary filter, which, together with our dictionary-based aligner, improves sense generation performance. Human analysis indicates that most errors are not inherent to the method, but rather arise from gaps in resources and noisy tools. Overall, our work represents a step forward for sense generation and its applications, providing a solid foundation for further developments.

## Acknowledgements

We thank Sevrin Robinson and Junhyeon Cho for their assistance. This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Alberta Machine Intelligence Institute (Amii).

## References

- [1] A. Oliver and S. Climent. “Automatic creation of WordNets from parallel corpora”. In: *Proc. LREC*. 2014, pp. 1112–1116.
- [2] F. Martelli, L. Procopio, E. Barba, and R. Navigli. “LexicoMatic: Automatic Creation of Multilingual Lexical-Semantic Dictionaries”. In: *Proc. IJCNLP-AAACL*. 2023, pp. 820–833.
- [3] C. Dyer, V. Chahuneau, and N. A. Smith. “A Simple, Fast, and Effective Reparameterization of IBM Model 2”. In: *Proc. NAACL-HLT*. 2013, pp. 644–648.
- [4] S. Neale. “A Survey on Automatically-Constructed WordNets and their Evaluation: Lexical and Word Embedding-based Approaches”. In: *Proc. LREC*. 2018.
- [5] R. Navigli and S. P. Ponzetto. “BabelNet: Building a Very Large Multilingual Semantic Network”. In: *Proc. ACL*. July 2010, pp. 216–225.
- [6] G. De Melo and G. Weikum. “Towards a universal wordnet by learning from combined evidence”. In: *Proc. CIKM*. 2009, pp. 513–522.
- [7] G. A. Miller, C. Leacock, R. I. Teng, and R. T. Bunker. “A semantic concordance”. In: *Proc. ARPA HLT*. 1993, pp. 303–308.
- [8] T. Pasini, A. Raganato, and R. Navigli. “XL-WSD: An Extra-Large and Cross-Lingual Evaluation Framework for Word Sense Disambiguation”. In: *Proc. AAAI* 35.15 (2021).
- [9] M. Jalili Sabet, P. Dufter, F. Yvon, and H. Schütze. “SimAlign: High Quality Word Alignments without Parallel Training Data using Static and Contextualized Embeddings”. In: *Findings of EMNLP*. 2020, pp. 1627–1643.
- [10] T. Ylonen. “Wiktextextract: Wiktionary as Machine-Readable Structured Data”. In: *Proc. LREC*. 2022, pp. 1317–1325.
- [11] D. Kamholz, J. Pool, and S. Colowick. “PanLex: Building a Resource for Panlingual Lexical Translation”. In: *Proc. LREC*. 2014, pp. 3145–3150.

## Appendix A. Experimental Setup

**Corpus.** We use the SemCor English sense-annotated corpus [7] as our source text, specifically the XL-WSD version [8]. This version is annotated with BabelNet synset IDs, which streamlines cross-lingual evaluation. Due to time constraints, we sample 20% of SemCor’s 37,176 sentences by taking every fifth sentence. The resulting sub-corpus, SC20, consists of 7,261 sentences containing 45,224 sense-annotated tokens. These annotations cover 12,251 BabelNet synsets, representing 9.7% of the noun and verb synsets in Princeton WordNet.

**Translation.** We utilize French as our development language and evaluate our method on Spanish, Chinese, and Urdu. For French, Chinese, and Spanish, we translate the corpus using `gpt-4o-2024-11-20` (the prompt used is provided in Figure B2). These translations are tokenized, lemmatized, and POS-tagged using `spaCy`. For Urdu, we obtain translations with `nllb-200-distilled-600M` and use `LughaatNLP` for the equivalent preprocessing.

**Alignment.** As the base aligner for DBAlign, we employ SimAlign [9] configured with BPE tokenization, embeddings from layer 8 of XLM-R, and the *itermax* matching algorithm.

**Dictionaries.** For our English-Spanish and English-French dictionaries, we take the union of all English-target translation pairs from WiktExtract [10] and PanLex [11]. From WiktExtract, we utilize all available pairs; from PanLex, we obtain all translations of any word appearing in any Princeton WordNet synset via the NLTK interface. For Chinese, we merge two freely available lexical resources: `CC-CEDICT` and `ECDICT`. The former provides domain-general vocabulary, while the latter offers substantially broader coverage. We merge all translation pairs from both sources, removing duplicates. Finally, for the English-Urdu dictionary, we combine dictionary files from `GoldenDict` with all English-Urdu pairs from WiktExtract. We then clean this concatenated dictionary by filtering out long phrases and extraneous entries in Hindi or Arabic.

## Appendix B. Prompts

Role	Content
System	You are a bilingual lexicon expert. Given a dictionary definition, produce the single word in {TARGET_LANGUAGE} that best matches this definition. Provide only the {TARGET_LANGUAGE} word without explanations.
User	{DICTIONARY_DEFINITION}

Figure B1. Prompt template for our LLM baselines.

Role	Content
System	You are an expert translator. Translate from {SOURCE_LANGUAGE} to {TARGET_LANGUAGE}. Provide only the translation without explanations.
User	{SOURCE_TEXT}

Figure B2. Prompt template for our translations.

## Appendix C. Parts of Speech

n	NOUN	PROPN	PRON	NUM	NN	PN	PP	GR	PD	D	CA	RP	KP	AP	REP
v	VERB	AUX	VB	AA	TA	KER	MUL								
a	ADJ	AKP	Q	OR	AD	KD									
r	ADV	ADP	PART	SCONJ	NEG	P	SE	QW	WALA	SC	I	FR	A		
x	CCONJ	INTJ	SYM	PUNCT	DET	X	SPACE	CC	SM	G	U	X	PM	EXP	INT

*Table C1.* Our mapping of POS tags from multiple tagsets to the basic POS categories. All tags mapped to *x* are considered to denote function words and are excluded from projection.

## Appendix D. Ablation

Aligner	Dict.	POS	Prec.	Cov.	HM
DBAlign	Used	Used	82.8	66.6	73.8
DBAlign	Used	Off	80.8	68.3	74.0
SimAlign	Used	Used	84.1	62.2	71.5
SimAlign	Used	Off	83.9	59.5	69.6
SimAlign	Off	Off	49.1	68.1	57.1

*Table D1.* Results for variants of **ExpandNet** on **SC20**, with Spanish as the target language. The Sense Precision, Synset Coverage, and their Harmonic Mean (HM) are reported for each method (in%). “Dict.” indicates whether a dictionary-based filter was used. Similarly, “POS” indicates whether the POS filter was used.