Semantic Interference: Polysemy Attenuates Cross-Lingual Embedding Alignment

Haokun Liu
Hypogenic AI Lab
haokun@hypogenic.ai

Abstract

1 Abstract

Multilingual Large Language Models (MLLMs) implicitly align representations across languages, creating a shared semantic space where translations map to similar vectors. However, this alignment assumes a one-to-one mapping between concepts, which is challenged by polysemy—where a single word form carries multiple distinct meanings. We investigate whether non-shared meanings in polysemous words act as a source of "semantic interference," pulling embeddings away from their translations. By analyzing English-French and English-Spanish pairs using `paraphrase-multilingual-MiniLM-L12-v2`, we find a statistically significant negative correlation ($ho = -0.125$, $p = 0.01$) between the number of non-shared senses and cosine similarity. While monosemous translations achieve high similarity (mean 0.82), polysemous words with non-shared meanings show measurable degradation. Crucially, however, we find that the shared semantic signal remains dominant: even with significant polysemy, translations remain far closer than random baselines (~ 0.37), suggesting robust, if imperfect, concept alignment in modern MLLMs.

2 Introduction

The promise of Multilingual Large Language Models (MLLMs) lies in their ability to map words and concepts from diverse languages into a shared, universal semantic space. If this mapping is perfect, a word in English and its translation in French should occupy the exact same point in the vector space, enabling zero-shot cross-lingual transfer Wu et al. [2019], Conneau et al. [2020]. This ideal, often termed "cross-lingual alignment" or "isomorphism," suggests that the geometry of meaning is universal. **But what happens when words have multiple meanings?** In reality, lexical structures rarely map one-to-one. The English word "bank" refers to both a financial institution and a river edge, while its French translation "banque" shares only the financial sense.

Understanding the fidelity of this shared space is critical. As MLLMs are increasingly deployed for cross-lingual retrieval and reasoning, we must understand the limits of their internal representations. If polysemy distorts embedding alignment, it introduces a fundamental noise floor for cross-lingual tasks. While recent work has confirmed that brains and language models converge on shared conceptual spaces Zada et al. [2025], and that linear alignments are possible Peng and Søgaard [2024], these studies often aggregate over large vocabularies, masking the granular impact of lexical ambiguity.

There is a gap in understanding the specific mechanics of this misalignment. Prior work acknowledges that non-isomorphism exists Zhang et al. [2019], typically attributing it to frequency differences or typological divergence. However, the systematic "pull" of non-shared meanings—the semantic interference caused by polysemy—has not been quantified. We lack a clear measure of how much a secondary, unshared meaning dilutes the primary semantic signal in the embedding space.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

We propose to quantify this phenomenon by correlating embedding similarity with a rigorous measure of "non-shared senses" derived from the Open Multilingual WordNet (OMW) Bond and Paik [2012]. We hypothesize that polysemy acts as a vector force: each distinct meaning pulls the embedding in a different direction. When a translation pair shares only one of these meanings, the non-shared senses act as noise, dragging the vectors apart. We test this on English-French and English-Spanish pairs using the `paraphrase-multilingual-MiniLM-L12-v2` model Reimers and Gurevych [2019].

Our experiments reveal a robust "semantic interference" effect. We find that while monosemous words (one shared meaning) achieve a high cosine similarity of extbf0.82, this drops as the number of non-shared senses increases. Specifically, we observe a statistically significant negative correlation of $ho = -\mathbf{0.125}$ ($p = 0.01$). This confirms that polysemy is a measurable drag on alignment. However, we also find the system is surprisingly resilient: even highly polysemous words maintain a similarity far above the random baseline of 0.37, indicating that the shared concept remains the dominant principal component of the representation.

Our contributions are:

- We extbfquantify the impact of polysemy on cross-lingual embedding alignment, introducing a "non-shared sense" metric.
- We extbfdemonstrate a statistically significant negative correlation between polysemy and embedding similarity in modern MLLMs.
- We extbfreveal that despite this interference, shared meanings remain the dominant signal, with embeddings showing resilience against semantic dilution.

3 Related Work

Implicit Cross-Lingual Alignment. Early work on Cross-Lingual Word Embeddings (CLWE) focused on explicit mapping techniques. Methods like MUSE Lample et al. [2018] learned linear transformations to align monolingual spaces, assuming structural isomorphism. However, the advent of large pretrained MLLMs revealed a surprising phenomenon: alignment emerges implicitly Wu et al. [2019]. Models like mBERT and XLM-R develop shared cross-lingual representations without explicit parallel data, likely due to shared subword tokens and structural similarities in language. We build on this understanding, treating the MLLM's internal representation as the shared space to analyze.

Concept Alignment. Recent studies have probed the nature of these shared spaces. Peng and Søgaard [2024] demonstrated high-quality linear alignment between concepts across languages in LLMs, using WordNet synsets as ground truth. Similarly, Zada et al. [2025] found convergence between human brain representations and LLM embeddings across languages, suggesting a universal conceptual geometry. While these works establish the existence of alignment, they often aggregate results, overlooking the granular impact of ambiguity. Our work zooms in on the specific instances where this alignment is challenged: polysemy.

Non-Isomorphism and Polysemy. The assumption of isomorphism—that languages share an identical geometric structure—has been challenged. Zhang et al. [2019] highlighted the issue of non-isomorphism, particularly for distant language pairs, and proposed iterative normalization to mitigate it. However, most prior work treats non-isomorphism as a global property or focuses on frequency effects. The specific role of polysemy as a local distorter of embedding similarity has received less attention. By quantifying the "pull" of non-shared senses, we provide a mechanism for understanding why certain concepts fail to align perfectly.

4 Methodology

We aim to quantify the relationship between semantic ambiguity (polysemy) and embedding similarity. Our approach combines lexical resources (dictionaries and WordNets) with dense vector representations from a state-of-the-art MLLM.

4.1 Data Construction

We construct a dataset of translation pairs annotated with sense counts. We use the MUSE dictionaries Lample et al. [2018] for English-French (`en-fr`) and English-Spanish (`en-es`) ground truth translations. To quantify polysemy, we align these pairs with the Open Multilingual WordNet (OMW) Bond and Paik [2012].

Filtering. We filter the dataset to include only pairs where both the source and target words are present in our OMW sample. This ensures reliable sense counts. To isolate semantic similarity from lexical overlap, we exclude pairs with identical spellings (e.g., "taxi" in `en` and `fr`), which would trivially have high similarity. This results in a high-quality dataset of 423 unique translation pairs.

4.2 Quantifying Polysemy

We define a metric for "semantic divergence" based on sense counts. Let $S(w)$ be the set of WordNet synsets (senses) for a word w. For a translation pair (w_{src}, w_{tgt}), the set of shared senses is $S_{shared} = S(w_{src}) \cap S(w_{tgt})$. The number of non-shared senses (NS) is defined as:

$$\mathrm{NS}(w_{src}, w_{tgt}) = |S(w_{src})| + |S(w_{tgt})| - 2|S_{shared}| \tag{1}$$

This metric captures the total number of meanings unique to either the source or target word. A value of 0 indicates perfect semantic equivalence (all senses are shared), while higher values indicate increasing divergence.

4.3 Model and Similarity Metric

We use `paraphrase-multilingual-MiniLM-L12-v2` (MINILM-L12) Reimers and Gurevych [2019] to generate embeddings. This model is distilled from XLM-R and optimized for semantic similarity, making it an ideal instrument for measuring cross-lingual alignment. For each word pair (w_{src}, w_{tgt}), we compute the cosine similarity of their embeddings:

$$\textsc{CosSim}(w_{src}, w_{tgt}) = \frac{\boldsymbol{v}_{src} \cdot \boldsymbol{v}_{tgt}}{\|\boldsymbol{v}_{src}\| \|\boldsymbol{v}_{tgt}\|} \tag{2}$$

where $\boldsymbol{v}$ is the embedding vector. We compare these similarities against a random baseline constructed by pairing source words with random target words.

5 Results

We investigate the alignment of translation pairs across two language pairs (`en-fr` and `en-es`) to test our hypothesis that non-shared senses act as a source of semantic interference.

5.1 Polysemy vs. Monosemy

We first categorize our translation pairs based on their sense overlap:

- **Monosemous-Shared**: Both words have exactly one sense, which is shared.
- **Polysemous-Full**: Both words have multiple senses, and all are shared.
- **Polysemous-Partial**: Words have one or more non-shared senses.

Table 1 summarizes the cosine similarities for these categories. Monosemous pairs exhibit high similarity (0.817), serving as a strong baseline for alignment. Surprisingly, fully shared polysemous pairs achieve even higher similarity (0.892), possibly due to their higher frequency or centrality in the semantic network. However, when non-shared senses are introduced (Polysemous-Partial), we observe a drop in similarity to 0.794. While this drop is statistically significant compared to the Polysemous-Full category, the values remain far above the random baseline (0.370).

5.2 Correlation Analysis

To quantify the interference effect, we calculate the Spearman correlation (ρ) between the number of non-shared senses (NS) and cosine similarity (COSSIM).

Table 1: Cosine Similarity by Polysemy Category. **Note:** Non-shared senses (Partial) reduce similarity compared to fully shared polysemy.

Category	**Mean Similarity**	**Std Dev**	**Count**
Monosemous-Shared	0.817	0.190	214
Polysemous-Full	**0.892**	0.109	10
Polysemous-Partial	0.794 ↓	0.177	49
Random Baseline	0.370	0.122	273

Global Trend. Across the entire dataset (n=423), we find a statistically significant negative correlation of $\rho = -0.125$ ($p = 0.010$). This confirms our hypothesis: each additional non-shared sense acts as a "pull," slightly degrading the alignment between the translation pair.

Language Variance. The effect is language-dependent. For English-Spanish pairs, the correlation is stronger ($\rho = -0.205$, $p = 0.003$), suggesting a higher sensitivity to polysemy or a greater divergence in sense structures between these languages. In contrast, English-French pairs show a weaker correlation ($\rho = -0.096$, $p = 0.155$), which may be attributed to the high lexical overlap and shared etymology between English and French.

5.3 Case Studies

To illustrate the "semantic pull," we examine specific pairs:

- **High Interference:** The pair 'shot' (`en`) vs 'coup' (`fr`) has 3 non-shared senses. Their similarity is notably low at **0.409**. While they share the sense of a "hit" or "blow," 'shot' also implies a projectile or photograph, meanings absent in 'coup' (which can mean a stroke, move, or political takeover).
- **Robust Alignment:** Conversely, 'substance' (`en`) and 'substance' (`fr`) share all major senses and achieve a similarity of **1.0**, demonstrating perfect alignment when conceptual structures match.

6 Discussion

Our results provide empirical evidence for "semantic interference" in multilingual embeddings. The negative correlation between non-shared senses and cosine similarity confirms that polysemy is not merely a linguistic curiosity but a quantifiable geometric force.

The "Pull" of Polysemy. We model this interference as a vector addition problem. If a word's embedding is the centroid of its meanings, then adding a non-shared meaning shifts this centroid away from the shared intersection. Our finding of $ho = -0.125$ validates this model: the more non-shared meanings, the further the shift.

Robustness of Shared Signal. Despite this interference, the most striking finding is the resilience of the shared signal. Even for words with significant polysemy (3+ non-shared senses), the similarity rarely drops below 0.5, remaining well above the random baseline of 0.37. This suggests that the shared core meaning—often the most frequent or dominant sense—anchors the embedding. The "pull" of secondary meanings is real, but it is insufficient to break the alignment.

Limitations. Our study relies on the Open Multilingual WordNet, which, while high-quality, has limited coverage compared to the full English WordNet. This restricts our sample size. Additionally, we use static embeddings from a Transformer model; contextualized embeddings (BERT/RoBERTa) likely resolve much of this ambiguity dynamically, a direction we leave for future work.

7 Conclusion

We investigated the impact of polysemy on cross-lingual word embedding alignment. By analyzing 423 translation pairs across English, French, and Spanish, we quantified a "semantic interference" effect: non-shared meanings systematically reduce the similarity between translations ($ho = -0.125$). However, we also found that this effect is bounded; modern MLLMs maintain robust alignment even in the face of ambiguity, suggesting a dominant shared conceptual core. These findings provide a granular view of the limits of implicit alignment and highlight the need for context-aware mechanisms in cross-lingual transfer.

References

Francis Bond and Kyonghee Paik. A survey of the OMW 1.0. In *Proceedings of the 6th Global WordNet Conference*, 2012.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*, 2020.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Word translation without parallel data. In *ICLR*, 2018.

Xinyu Peng and Anders Søgaard. Concept space alignment in multilingual LLMs. In *Proceedings of ACL*, 2024.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of EMNLP*, 2019.

Shijie Wu, Alexis Conneau, Haotian Ha, Douglas Oard, and David Yarowsky. Emerging cross-lingual structure in pretrained language models. In *Proceedings of ACL*, 2019.

Zaid Zada, Ariel Goldstein, Samuel A Nastase, and Uri Hasson. Brains and language models converge on a shared conceptual space across different languages. *Nature Human Behaviour*, 2025.

Mozhi Zhang, Koichiro Yoshino, Sakriani Sakti, and Satoshi Nakamura. Are girls neko or shōjo? cross-lingual alignment of non-isomorphic embeddings with iterative normalization. In *Proceedings of ACL*, 2019.

A Appendix

A.1 Dataset Statistics

Our dataset was constructed using the intersection of MUSE dictionaries and the Open Multilingual WordNet (OMW). Table 2 provides a breakdown of the number of pairs per language and polysemy category.

Table 2: Detailed Dataset Statistics

Category	**EN-FR**	**EN-ES**	**Total**
Monosemous-Shared	142	72	214
Polysemous-Full	6	4	10
Polysemous-Partial	31	18	49

A.2 Implementation Details

We use the ‘sentence-transformers‘ library to load the ‘paraphrase-multilingual-MiniLM-L12-v2‘ model. Embeddings are extracted from the ‘[CLS]‘ token for single words. The code is available at `https://github.com/hypogenic/similar-embeddings-nlp`.