Cognitive Landscapes of Language: Investigating Polysemy Representation in English, German, and Multilingual BERT Models

Harshini Raju[1], **Shambhavi Seth**[1], **Varshitha Reddy Medarametla**[1], **Rishabh Raj**[1]
[1]Department of Computer Science, New York University
{hr2547, ss17936, vm2663, rr4574}@nyu.edu

Abstract

The human brain's linguistic and cognitive abilities have enabled remarkable communication and survival mechanisms. This paper investigates the phenomenon of polysemy—words with multiple meanings—within English and German using monolingual and multilingual BERT models. By analyzing embeddings for polysemous words, we seek to understand the interaction of language, cognition, and computational modeling.

Introduction

The human brain, with its immense cognitive capabilities, underpins survival, communication, and thought. The dynamic relationship between language and cognition has fascinated scholars across disciplines, with communication playing a key role in cultural evolution and historical development.

In the 20th century, Noam Chomsky challenged behaviorist views, arguing for innate linguistic structures in *Syntactic Structures* (1957). Building on this, researchers like Boroditsky (Boroditsky (2011)) demonstrated how native language grammar influences perception, while Athanasopoulos et al. (2015) highlighted how bilinguals' cognition adapts to the language in use.

Polysemy, where a word has multiple related meanings (Pustejovsky, 2021), is a key linguistic challenge. Deane (Deane (1988)) described it as a cognitive relevance effect, aligning with Saussure's theory of linguistic signs as a combination of form and concept. Early models like Word2Vec and GloVe failed to address polysemy, assigning static embeddings to words. Contextualized models like BERT overcome this by generating embeddings that adapt to usage, but polysemy remains a challenge influenced by linguistic structure and cultural context.

This study examines polysemy in English and German, two Indo-European languages with shared roots but distinct grammatical systems. By analyzing contextualized embeddings from monolingual and multilingual BERT models, it explores how language and architecture influence polysemy representation, contributing to our understanding of computational language modeling and cognition.

Related Work

The development of contextualized language models like BERT (Devlin, Chang, Lee, and Toutanova (2019)) has significantly advanced Natural Language Processing (NLP). Early models such as Word2Vec (Mikolov, Chen, Corrado, and Dean (2013)) and GloVe (Pennington, Socher, and Manning (2014)) mapped words to dense vectors based on semantic relationships but failed to address polysemy due to static embeddings. BERT overcame this limitation by generating dynamic embeddings that adapt to word context, excelling in tasks such as language inference and word sense disambiguation (Huang, Sun, Qiu, and Huang (2019)).

Evolution of Contextualized Embeddings

Word embedding methods like Word2Vec and GloVe focused on capturing local and global co-occurrence patterns, respectively, but assigned static representations, making them inadequate for handling polysemy. The introduction of contextualized models like ELMo (Peters et al. (2018)) and BERT (Devlin et al. (2019)) enabled dynamic embeddings that adjust to context, marking a major shift in NLP. Studies such as Garí Soler and Apidianaki (2021) demonstrated BERT's effectiveness in distinguishing polysemous meanings through context, while refinements like RoBERTa (Liu et al. (2019)) enhanced contextual variation capture.

Lexical Ambiguity and Cognitive Modeling

Studies (Huang et al. (2019)) highlight BERT's ability to manage lexical ambiguity, particularly in polysemous words tied to perception and cultural factors (Whitt (2009)). These findings align with cognitive theories suggesting a continuum between polysemy and homonymy, emphasizing the role of context in disambiguation. Such insights are vital for bilingual contexts, where linguistic systems interact dynamically.

English-German Comparisons in Polysemy

English and German provide an ideal comparison for polysemy analysis due to their shared linguistic roots and distinct grammatical systems (Whitt (2009)). For example, while English relies on word order, German uses a case system, influencing the representation and interpretation of polysemous words. Cross-linguistic studies offer valuable perspectives on how contextualized models handle linguistic variation.

Multilingual Models and Alignment

Multilingual BERT (mBERT) captures cross-lingual embeddings but faces challenges in aligning polysemous meanings across languages (Liu, Yuan, and Schmitt (2020)). This discrepancy is partly due to challenges in aligning polysemous word meanings across different linguistic and cultural contexts. Contextual alignment strategies (Cao, Kitaev, and Klein

(2020)) improve the ability of multilingual models to represent polysemy, enabling better cross-lingual transfer.

Experimental Basis for Current Study and Key Insights

This study builds on Athanasopoulos et al. (2015) and Garí Soler and Apidianaki (2021) to evaluate how monolingual and multilingual BERT models represent polysemy in English and German. While BERT excels at capturing contextual nuances, aligning multilingual embeddings to match monolingual performance remains challenging. Fine-tuning and alignment strategies offer potential solutions, enhancing both model performance and understanding of bilingual cognitive processes.

Dataset

Dataset Preparation

To ensure a meaningful and equal comparison between the English and German data in monolingual models and the multilingual BERT, the dataset of 100 polysemous words in each language. Each word is accompanied by two pair of sentences, with each pair representing a distinct meaning of the word. This dataset design allows us to observe how the models embed the same word differently across varying contexts. For example, for the word bank the sentences in English are :

1. **Bank**

   (a) She went to the **bank** to deposit her paycheck.

   (b) The teller at the **bank** helped her open a new savings account.

   (c) The river overflowed its **bank** after the heavy rain.

   (d) Children were playing on the grassy **bank** of the lake.

In the first pair of sentences, the word *Bank* refers to a *financial institution*, while in the second pair of sentences, it refers to the *river bank*. Similarly, in German:

1. **Bank**

   (a) Sie ging zur **Bank**, um ihren Gehaltscheck einzuzahlen.

   (b) Der **Bank**angestellte half ihr, ein neues Sparkonto zu eröffnen.

   (c) Der Fluss trat nach dem starken Regen über sein **Ufer**.

   (d) Kinder spielten am grasbewachsenen **Ufer** des Sees.

In the first pair of sentences, *Bank* refers to a financial institution (*Finanzinstitut*), whereas in the second pair of sentences, it denotes a riverbank (*Flussufer*).

The inclusion of both English and German sentences ensures that language-specific nuances, such as differences in polysemy representation and contextual interpretation, are captured effectively.

Cross-linguistic Balance

The dataset balances semantic ranges and sentence structures in English and German, enabling comparisons while accounting for syntactic and morphological differences.

Though both Indo-European, English relies on word order, while German's case system shapes word usage (Whitt, 2009).

Related Datasets

Our dataset builds upon and complements existing resources such as:

1. **Homonymy Dataset by Marcos and et al. (2021):** This dataset focuses on lexical ambiguity, particularly distinguishing polysemy from homonymy.

2. **RAW-C Corpus by Trott and et al. (2021)** This dataset emphasizes word-sense disambiguation across different corpora, we drew from its structured approach to represent sentence-level context and meaning.

The dataset enables analysis of polysemy in English and German, supporting comparisons of monolingual BERTs and mBERT in handling lexical ambiguity. It balances grammatical roles, semantic ranges, and linguistic structures for robust evaluation.

Experimental Setup

Models

This study uses three pre-trained BERT models—English BERT, German BERT, and Multilingual BERT (mBERT)—to analyze contextualized embeddings for polysemy.

1. **English BERT:** BERT (Bidirectional Encoder Representations from Transformers), introduced by Devlin et al. (2019), is pre-trained on large English corpora (BookCorpus, Wikipedia) and relies on bidirectional attention to capture nuanced word meanings in context. Key features include:

   BERT's architecture consists of:

   (a) Transformer Layers: A stack of 12 layers for the base model (24 for the large model).

   (b) Attention Heads: Each layer includes 12 attention heads for multi-head self-attention (16 for the large model).

   (c) Hidden Units: Each layer has 768 hidden units (1024 for the large model).

   The pretraining objectives for BERT include masked language modeling (MLM) and next sentence prediction (NSP), which collectively allow the model to capture deep semantic relationships.

2. **German BERT:** Adapted for German, this model uses the same architecture as BERT but is pre-trained on German-specific datasets (e.g., Open Legal Data, German Wikipedia) for deeper understanding of German's syntactic and morphological features like case and compounding(Chan, Schweter, and Möller (2020)).

3. **Multilingual BERT (mBERT):** Pre-trained on 104 languages, mBERT uses a shared vocabulary of 110,000 tokens to capture cross-lingual semantics. While it lacks explicit cross-lingual objectives, research shows it effectively aligns multilingual semantic spaces, enabling polysemy analysis across languages (Pires, Schlinger, and Garrette (2019)).

Comparison of Models

The choice of English BERT, German BERT, and mBERT allows us to explore polysemy from three perspectives:

1. **Monolingual Performance**: English BERT and German BERT provide insights into language-specific handling of polysemous words.

2. **Cross-Lingual Alignment**: mBERT facilitates comparisons of polysemous word representations between English and German within a shared embedding space.

By analyzing these three models—English BERT, German BERT, and mBERT—we aim to understand the nuanced handling of polysemous words across monolingual and multilingual contexts. The following sections provide detailed insights into their architecture and comparative performance.

Contextual Embeddings in BERT Models

Contextual embeddings represent words dynamically, with their meanings shaped by surrounding words. For example, the word bat in "He swung the bat with all his might" generates an embedding distinct from bat in "A bat flew out of the cave." This dynamic representation is achieved through self-attention mechanisms in transformer-based architectures, enabling the model to weigh the importance of context words differently for each occurrence of a target word.

Differences Across Models:

1. **Vocabulary:**
   (a) *English BERT* and *German BERT* use vocabularies tailored to their respective languages, allowing them to represent words with greater granularity in language-specific contexts.
   (b) *mBERT*, in contrast, uses a shared vocabulary across 104 languages. While this facilitates cross-lingual alignment, it can dilute the model's ability to capture fine-grained language-specific details, particularly for polysemous words with culturally or linguistically unique meanings.

2. **Pretraining Data:**
   (a) *English BERT* is pre-trained on the *English Wikipedia* and *BookCorpus*, which ensures a robust understanding of English-specific linguistic patterns.
   (b) *German BERT* is pre-trained on German corpora, including *Open Legal Data* and the *German Wikipedia*, enabling it to learn nuances such as compound words and a case-based grammatical system.
   (c) *mBERT* is pre-trained on Wikipedia texts from multiple languages. This broad scope enables it to generalize across languages but makes it less specialized in individual languages like English or German (Liu et al. (2020)).

3. **Embedding Space:**
   (a) Monolingual models (*English BERT, German BERT*) represent word embeddings within language-specific spaces. As a result, they may produce embeddings that are more sensitive to subtleties in polysemous words unique to their respective languages.
   (b) *mBERT* creates a shared embedding space for multiple languages, which can align semantically similar words across languages but may struggle with precise disambiguation for polysemy in individual languages.

Differences in Contextual Embedding Dynamics

1. **Handling of Polysemy:**
   (a) Monolingual models like *English BERT* and *German BERT* excel at distinguishing polysemous word meanings because their vocabularies and pretraining are fine-tuned to the nuances of the language.
   (b) *mBERT*, while capable of cross-lingual alignment, may exhibit less granularity in distinguishing polysemous meanings due to its shared vocabulary.

2. **Cross-Lingual Alignment:**
   (a) *mBERT*'s shared embedding space enables direct comparison of polysemous word representations across languages. For instance, *bank* in English and *Bank* in German may cluster closely in *mBERT*'s embedding space, reflecting their semantic similarity.
   (b) However, this alignment can come at the cost of reduced precision for language-specific distinctions.

3. **Adaptability to Context:**
   (a) All models use self-attention to generate embeddings that adapt to the surrounding context. However, monolingual models are better at leveraging language-specific grammatical patterns, while *mBERT*'s adaptability is influenced by its multilingual corpus.

Significance of Model Comparison

By leveraging these three models, this study examines:

1. **Monolingual Performance**: How well English BERT and German BERT represent polysemous words within their respective languages.

2. **Cross-Lingual Performance**: How mBERT aligns and disambiguates polysemous words across English and German.

3. **Model Limitations**: Where monolingual and multilingual models struggle, particularly in handling highly context-dependent or linguistically unique polysemy.

Pipeline

This pipeline explores semantic representations of polysemous words across German and English using transformer-based models. It employs the pre-trained German BERT model (`dbmdz/bert-base-german-uncased`) and English BERT model (`bert-base-uncased`) with `output_hidden_states=True` to extract hidden states from all transformer layers. These embeddings undergo preprocessing, dimensionality reduction, and clustering to reveal semantic patterns. Tokenizers specific to each model preprocess sentences in their respective languages.

1. **Sentence Preprocessing:** Sentences were tokenized into subword tokens using language-specific tokenizers. Target words were carefully identified in tokenized outputs, even when subword splitting occurred, ensuring accurate embedding alignment for cross-lingual comparisons.

2. **Embedding Extraction:** Word embeddings for target words were extracted from all hidden layers of the BERT models. For words occurring multiple times in a sentence, their embeddings were averaged to capture contextual meaning. This process was repeated for four sentences per word in both languages, maintaining consistency in handling sentences with the same or different senses of each word.

3. **Normalization Techniques:** To enhance the embeddings' comparability and remove potential biases, we applied a series of normalization steps:
   (a) **Z-Score Normalization:** Standardized embeddings to a mean of zero and unit variance, equalizing scales across dimensions to prevent dominant features from skewing similarity or clustering results.
   (b) **Quantile Normalization:** Adjusted embedding values to follow a Gaussian distribution, reducing outliers and ensuring consistency in similarity computations.
   (c) **Whitening:** Removed correlations and scaled dimensions to equal variance, transforming embeddings into an uncorrelated space for improved interpretability and reduced redundancy.
   (d) **Unit Norm Normalization:** Scaled embeddings to unit length, focusing on vector direction to ensure similarity measures like cosine similarity reflect semantic relationships rather than magnitude differences.

4. **Clustering and Evaluation:** To analyze contextual meanings, we applied t-SNE which gives us clusters of same and different sense words normalized embeddings while also reducing the dimensions to 2 for easy visualization. This separates the normalized word embeddings for different senses of the same word and allows for a nuanced comparison of how different models and languages represent word senses, potentially revealing insights into cross-lingual semantic relationships and the effectiveness of multilingual models in capturing sense distinctions.

5. **Semantic Similarity Analysis:** To quantify the similarity between embeddings, we computed cosine similarity. This metric reveals how close two embeddings are in the vector space, indicating semantic proximity. Three key similarity metrics were calculated:
   (a) **Same-Sense Similarity (within-language)**: Compared embeddings from sentences reflecting the same sense of a word, such as two sentences for "Bank" as a financial institution in German or English.
   (b) **Different-Sense Similarity (within-language)**: Compared embeddings from sentences reflecting different senses of a word, such as one sentence for "Bank" as a financial institution and another for "Bank" as a riverbank in German or English.
   (c) **Cross-Language Similarity**: Assessed similarities between German and English embeddings for the same and different senses of polysemous words to evaluate cross-lingual alignment.

6. **Layer-Wise Analysis:** The evolution of semantic representations across transformer layers was analyzed to study how polysemous word senses were differentiated at varying depths within the models. Layer-wise similarity trends were examined to understand the progressive refinement of semantic representations. Cross-language comparisons were also performed to evaluate how effectively the semantic spaces of the German and English models aligned.

7. **Dimensionality Reduction and Visualization:** To visualize the clustering of embeddings, t-SNE was applied to reduce the high-dimensional embeddings from the final transformer layers to two dimensions. These visualizations highlighted the clustering patterns of sentences based on the sense of the target word within and across languages. Sentences representing the same sense in German and English were expected to form overlapping clusters, while different senses formed distinct clusters.

8. **Comparative Evaluation:** The comparative analysis focused on assessing each model's ability to:
   (a) Differentiate between the senses of polysemous words within its language.
   (b) Align semantic representations for polysemous words across German and English.

Results

Word Sense Embeddings Analysis

The analysis of word sense embeddings across monolingual and multilingual BERT models revealed distinct patterns in how different senses are represented in the embedding space. The plot in Figure 1 is a clustering representation of 768 dimensional vectors in 2 dimensions using t-SNE. We can see clustering of model specific embeddings and can see the semantic group of pairs of sentences as closer together and far from its polysemic counterpart. More details in Figure 3

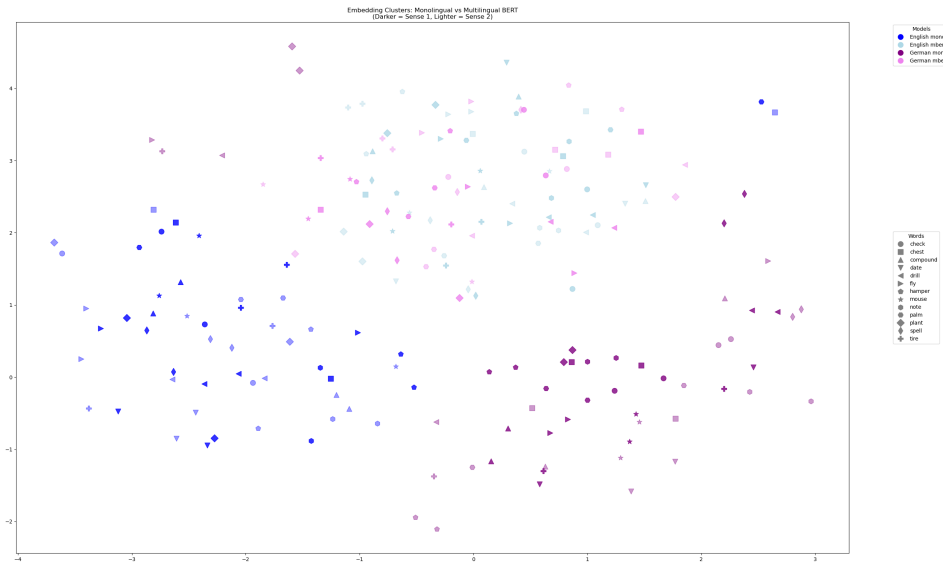

Figure 1: Embedding clusters visualization showing the distribution of word senses across monolingual and multilingual BERT models. Darker points represent Sense 1, while lighter points represent Sense 2.

Model Performance Analysis

Monolingual BERT Performance: This model demonstrates high initial cosine similarity (∼0.95) for same-sense pairs (1-2), with a gradual decline across layers. The model shows clear differentiation between same-sense and different-sense pairs, with different-sense similarity (1-3) dropping significantly to (∼0.45) by layer 12.

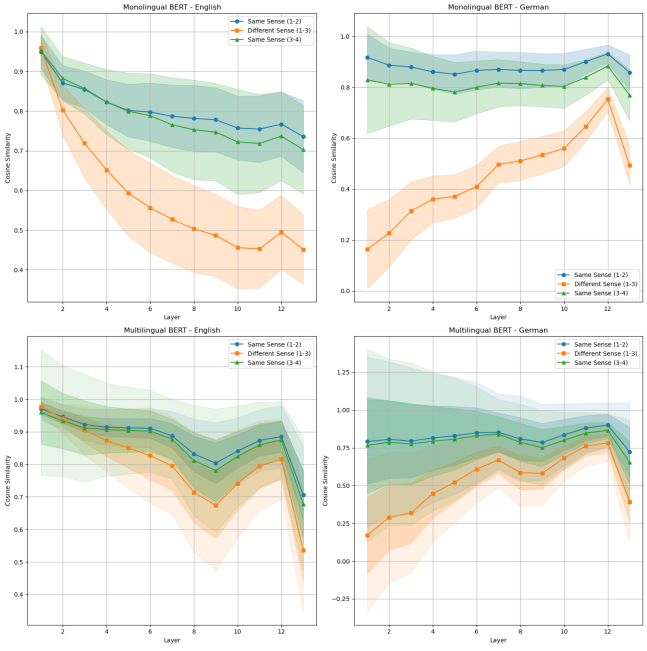

Figure 2: Cosine similarity across layers for monolingual and multilingual BERT models in English and German. Shaded areas represent standard deviation.

German monolingual BERT: It exhibits more stable same-sense similarity (∼0.9) across layers, with better maintenance of sense distinctions and higher baseline similarity for different senses.

Multilingual BERT: But with BERT multilingual plots, the different sense is closer to the same sense similarity score, this highlights that monolingual BERT is better at capturing the polysemic nuance of language, but this differentiation is somewhat blurred in multilingual. It still is able to score them lower than the same sense pairs across all layers.

Statistical Significance

The standard deviation analysis, depicted by the shaded areas in Figure 2, reveals several key insights: early layer representations exhibit higher confidence, while deeper layers show increasing uncertainty in the English BERT model. In contrast, German BERT demonstrates more stable representations compared to its English counterpart. Furthermore, multilingual models display greater variability, particularly in their ability to map cross-lingual senses.

Word-Specific Analysis

Analysis revealed the strongest sense separation for:

1. **plant**: biological vs. industrial contexts
2. **hamper**: container vs. obstruct meanings
3. **palm**: tree vs. hand meanings

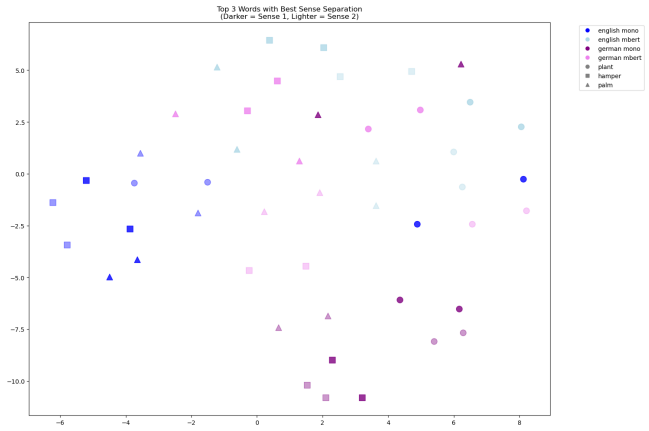

Figure 3: Zoomed in version of the cluster(highlighting 3 words) showing the strongest sense separation across models. The visualization demonstrates clear clustering patterns for 'plant', 'hamper', and 'palm'.

Conclusion

This study reveals the strengths and limitations of contextualized embeddings in the representation of polysemy. Monolingual models excel in language-specific nuances, while mBERT offers cross-lingual alignment but with reduced granularity. These findings contribute to the understanding of polysemy in NLP and cognitive modeling. The results demonstrated that both the German and English BERT models effectively differentiated between word senses within their respective languages. Cross-lingual alignment between German and English embeddings showed strong semantic correspondence for the same sense of polysemous words, validating the effectiveness of the bilingual pipeline. Homonomy still poses a big challenge for these contextualized models. As also evident from the clustering as well as the similarity plots.

References

Athanasopoulos, P., Bylund, E., Montero-Melis, G., Damjanovic, L., Schartner, A., Kibbe, A., ... Thierry, G. (2015). Two languages, two minds: Flexible cognitive processing driven by language of operation. *Psychological Science*, *26*(4), 518–526.

Boroditsky, L. (2011). How language shapes thought. *Scientific American*, *304*(2), 62–65.

Cao, S., Kitaev, N., & Klein, D. (2020). Multilingual alignment of contextual word representations. *arXiv preprint arXiv:2002.03518*.

Chan, B., Schweter, S., & Möller, T. (2020). German's next language model. In *Proceedings of the 28th international conference on computational linguistics* (pp. 6788–6796). Barcelona, Spain (Online): International Committee on Computational Linguistics. Retrieved from https://aclanthology.org/2020.coling-main.598 doi: 10.18653/v1/2020.coling-main.598

Deane, P. D. (1988). Polysemy and cognition. *Lingua*, *75*(4), 325–361. doi: https://doi.org/10.1016/0024-3841(88)90009-5

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacl-hlt 2019* (pp. 4171–4186).

Garí Soler, A., & Apidianaki, M. (2021). Let's play monopoly: Bert can reveal words' polysemy level and partitionability into senses. *Findings of ACL 2021*.

Huang, L., Sun, C., Qiu, X., & Huang, X. (2019). Glossbert: Bert for word sense disambiguation with gloss knowledge. In *Proceedings of emnlp-ijcnlp 2019* (pp. 3509–3514). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1355 doi: 10.18653/v1/D19-1355

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Y., Yuan, H., & Schmitt, N. (2020). On the multilingual contextual embeddings in mbert and their cross-lingual alignment. In *Proceedings of the emnlp workshop.*

Marcos, J., & et al. (2021). Homonymy dataset for lexical ambiguity. *ACL 2021*. Retrieved from https://github.com/example/homonymy-dataset

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of naacl-hlt 2018* (pp. 2227–2237).

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual bert? In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 4996–5001). Florence, Italy: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P19-1493 doi: 10.18653/v1/P19-1493

Trott, S., & et al. (2021). Raw-c corpus: Dataset for word-sense disambiguation. *GitHub Repository*. Retrieved from https://github.com/example/raw-c-corpus

Whitt, R. J. (2009). Evidentiality, polysemy, and the verbs of perception in english and german. *Annual Review of Cognitive Linguistics*, *7*(1), 120–150.