Universiteit Leiden
The Netherlands

Data Science and Artificial Intelligence

Semantic Priming: Comparing LLMs, Human Feature Norms and Associative Strength

Figen Ulusal

First and second supervisor:
Tom Heyman and Evert van Nieuwenburg

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

01/07/2025

Abstract

This thesis explores how different measures of semantic similarity predict semantic priming effects in a lexical decision task. Semantic priming refers to the phenomenon where a word is recognized faster when preceded by a semantically related word, reflecting how concepts are structured in the mental lexicon. The study compares three types of similarity: large language model (LLM)-based embeddings, human-generated feature norms, and associative similarity. Using the English subset of the SPAML dataset, Study 1 evaluates the predictive power of LLM-based and human-based feature overlap. Results show that while all predictors significantly correlate with priming effects, LLM-based similarity performs best. Study 2 adds associative similarity to the comparison and finds that it does not significantly predict priming nor improve model performance. Across both studies, LLM-based similarity consistently outperforms other measures, though feature norms offer some complementary value. These findings emphasize the effectiveness of LLMs in modeling meaning and suggest that distributional representations can reflect aspects of human lexical processing. This contributes to understanding how computational models approximate human semantic knowledge and supports the use of LLMs in psycholinguistic research.

Contents

1 General Introduction

Large language models (LLMs) can now perform a range of linguistic tasks with human-like fluency. But do they capture the way humans mentally organize language and can they do it better than human intuition themselves? The human mind does not store words like entries in a dictionary. Instead, words are interconnected in a mental network, where meaning arises from links between related concepts. One of the earliest theories about how meaning is stored in the mind was proposed by Collins and Quillian [CQ69]. They proposed that semantic memory is structured as a hierarchical network of nodes and links, with general properties stored at higher-level nodes and are inherited by more specific concepts. An illustration of this semantic memory model is shown in Figure 1. In their framework, when thinking about one concept (e.g., *bird*), activates connected concepts (e.g.,*canary* or *animal*), making them easier to retrieve from the mind. Collins and Quillian's model laid the theoretical groundwork for exploring how connections between concepts might influence the speed of language processing. Building on this, Meyer and Schvaneveldt [MS71] demonstrated that people respond faster to a word when it is follows a related word, which is an effect now known as semantic priming. For example, individuals recognize the word *dog* more quickly when it is preceded by the word *cat* than when it is preceded by an unrelated word like *bus*. This thesis explores different measures of predicting semantic priming, aiming to determine whether semantic relationships can be modeled computationally by both LLM-based measures and human-based measures.

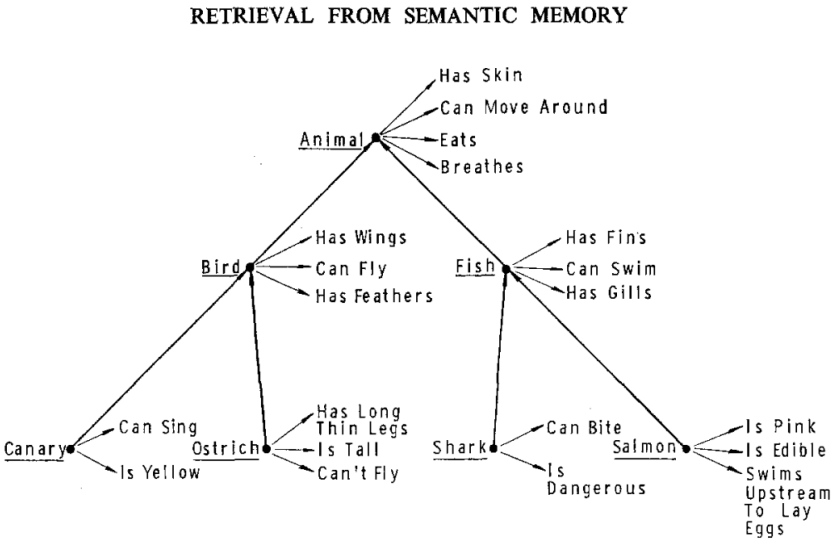


Figure 1: Illustration of the hypothetical memory structure for a three-level hierarchy. Adapted from Collins and Quillian [CQ69, Fig. 1].

1.1 Thesis overview

This thesis is divided into two studies that evaluate how different similarity measures predict semantic priming effects. Study 1 compares predictions based on human- and LLM-based feature overlap, using the full English subset of the SPAML dataset. Study 2 extends this comparison by adding a third predictor: associative strength from the SWOW dataset. Each study begins with an introduction to the relevant theoretical background, followed by a detailed explanation of the datasets and analysis methods. The results are then presented and interpreted in dedicated discussion and conclusion sections. The thesis concludes with a general discussion that reflects on the findings, addresses limitations, and outlines directions for future research.

2 Study 1: Predicting Semantic Priming from Human- and LLM-Based Feature Overlap

2.1 Introduction

This study compares human-based and LLM-based feature overlap in predicting semantic priming. To understand how these two measures predict semantic priming, this section discusses key theories on how the mind is organized and accesses word meanings (Section 2.1.1). It then introduces theoretical models of semantic memory (Section 2.1.2) and describes how semantic priming is measured experimentally (Section 2.1.3). After that large-scale datasets used to investigate semantic priming (Section 2.1.4) are discussed, this section will elaborate on the approaches used to predicting semantic priming (Section 2.1.5). The section ends with the research questions addressed by this study.

2.1.1 Semantic Priming and the Mental Lexicon

In psycholinguistics, semantic priming is used to study the structure of the mental lexicon. The mental lexicon was traditionally defined as a dictionary of all lexical knowledge that a person possesses [SW08]. However, this view has been criticized for treating lexical knowledge as passively stored in the main, which fails to explain how words are accessed across different sensory domains, such as listening and reading, and being unclear about whether people rely on a single or multiple lexicons. As a result, the mental lexicon is now seen as a dynamic and distributed system that enables the ability for lexical activity. The mental lexicon enables individuals to efficiently access and use appropriate words during everyday language use, even in new or changing contexts. Semantic priming is a common method for studying the mental lexicon, as it shows how related the activation of one word can influence the accessibility of another during real-time language processing[MKB17, BVM19].

2.1.2 Theoretical Models of Semantic Memory

Understanding how semantic priming occurs requires the understanding of how concepts are structured in the memory. To explain this structure, researchers have proposed models that represent concepts as interconnected nodes. One influential early model, as briefly introduced in Section 1, was by Collins and Quillian [CQ69] proposed a hierarchical model explaining the

organization of concepts in our mind. Due to its rigid structure, it struggled to account for semantic relationships that do not fit into the rigid categories, such as concepts related by function or experience instead of taxonomy. To address these limitations, Collins and Loftus [CL75] proposed a more flexible spreading activation model, concepts are still being represented as nodes in a network, but are now connected by links that vary in length, strength and direction. This is illustrated in Figure 2. These properties determine how activation flows through the network: shorter links allow faster spreading, stronger links transmit more activation and directionality accounts for asymmetries in activation (e.g., *fire* may activate *smoke* more strongly than the reverse). When one concept is retrieved, activation automatically spreads to nearby and strongly connected nodes, lowering their retrieval thresholds and making them easier to access. The closer and more strongly linked a concept is, the faster and more likely it is to be accessed.

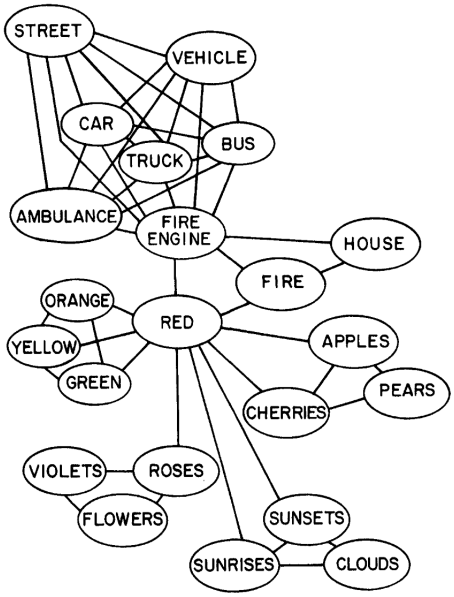


Figure 2: Schematic representation of conceptual relatedness in a typical fragment of semantic memory. Shorter lines indicate greater relatedness. Adapted from Collins and Loftus [CL75, Fig. 1].

2.1.3 Measuring Semantic Priming

Based on this spreading activation model, the closer two concepts are in a semantic network, the faster one should be recognized after the other [CL75]. This forms the basis of semantic priming: people recognize a word more quickly when it follows a related word than when it follows an unrelated one. Semantic priming is typically measured by comparing response times across conditions, with the assumption that related cues, lower the activation threshold to recognize a target, making response times a useful indicator of how closely two concepts are linked in memory. The most common methods to measure semantic priming is through reaction time tasks, such as a naming task and the lexical decision task. In a naming task, participant are instructed to pronounce the target word aloud (e.g., *table*) as quickly and accurately as possible, after seeing a cue word that is either semantically related (e.g., *chair*) or unrelated (e.g., *watch*) [HBN+13]. In lexical decision tasks, in contrast, participants must decide as quickly as possible whether a string of letters forms

a real word.This method was first demonstrated by Meyer and Schvaneveldt [MS71], who showed that people recognize a word like *butter* more quickly when it is preceded by a related cue like *bread* than by an unrelated cue. Their results provided the first direct behavioral evidence for spreading activation in semantic memory. Later, work by Neely [Nee77] showed that priming effects can occur automatically, consistent with the spreading activation theory, but also strategically, like expectancy generation. Strategic effects were especially visible when participants had more time to process the cue, showing that the timing of stimulus presentation plays an important role in whether priming reflects automatic or strategic processing. To reduce such strategies and better isolate automatic processes, researchers have developed alternative task designs, including the continuous lexical decision task [BCC+21]. In the continuous lexical decision task, participants make a word/non-word judgment for every item in a fast-moving sequence, including both cues and targets. Because they are not explicitly told which words form a pair, it becomes harder to anticipate upcoming targets or rely on expectations, reducing the impact of strategic processing. These lexical decision tasks have since been used not only to study the existence of semantic priming across word pairs, but also to examine differences in priming strength between word pairs. This enabled the creation of datasets aimed at predicting priming effects across many word pairs.

2.1.4 Large-Scale Datasets

Although lexical decision tasks have proven useful for measuring semantic priming, most datasets based on them are limited in size and lack reliability. They often include a small number of hand-selected word pairs, which restricts their use for large-scale or generalizable analyses. To address these limitations, the Semantic Priming Project (SPP) was the first large-scale effort to collect semantic priming data across a wide range of stimuli and conditions [HBN+13]. The dataset includes over 1,600 cue–target combinations and responses from more than 750 participants. Importantly, it manipulates two important variables, namely the task type (naming tasks and lexical decision tasks) and the stimulus onset asynchrony (SOA), which is the amount of time between presenting the cue and target. By including both naming and lexical decision tasks, and including both shorter SOA's (200 ms) and longer SOA's (1,200 ms), SPP reflects both automatic priming mechanisms and and more strategic processes. Despite these strengths, SPP has three main limitations [BCC+21]. First, the result on item-level showed low reliability, making it difficult to reliably predict priming effects. Relatedly, the sample size per item was too small to reliably predict predict priming effects, indicating the need for a larger dataset. Lastly, SPP only contains English data.

The SPAML project (Semantic Priming Across Many Languages) was designed to address these limitations. SPAML substantially increases statistical reliability by implementing adaptive sampling procedures and collecting responses from over 25,000 participants across 19 languages [BCC+21]. Unlike SPP, where each item was seen by only a handful of participants, SPAML ensures that each word pair is evaluated across a larger and more diverse sample. Furthermore, by including multiple languages in a standardized task, the continuous lexical decision task, SPAML enables cross-linguistic comparisons while reducing strategic processing. This makes the SPAML dataset the largest semantic priming dataset to date, offering a reliable and broadly applicable foundation for investigating semantic priming.

2.1.5 Predicting Semantic Priming

A common approach to predicting semantic priming is based on calculating semantic similarity. Semantic similarity refers to the degree to which two concepts are related in meaning and can be defined in multiple ways [BCC+21]. One such approach is feature-based similarity, which is commonly referred to as feature overlap [MCSM05]. These features, such as *has fur, is a pet* and *meows* for the concept *cat*, are typically derived from human-generated feature norms. The more features two concepts share, the more semantically similar they are considered to be and the stronger the expected priming effect is. This approach assumes that two concepts share many semantic features, activating one concept (the cue) will partially activate the other concept (the target), facilitating individuals to recognize the target word faster. Feature overlap, the degree of features that two concepts share, can thus serve as a predictor of priming effects. To measure feature overlap, researches often rely on semantic feature norms. According to McRae et al. [MCSM05], when a cue and target word share many features, thus have more feature overlap, the target word is recognized more quickly after the cue. This is due to overlapping activations in the mental representations.
This idea has been further explored by Buchanan et al. [BVM19] by testing cosine similarity scores derived from semantic feature norms against semantic priming from the Semantic Priming Project, discussed in section 2.1.4. Cosine similarity is a commonly used metric that measures the similarity between two non-zero vectors, in this case semantic feature norms, The resulting similarity score, as defined in equation 1, reflects the amount of feature overlap. A score of 1 indicates a complete overlap and a score of 0 indicates no overlap.

$$\text{similarity}(\mathbf{A}, \mathbf{B}) = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1}$$

Results from this study showed that correlations between feature-based cosine similarity and priming effects were generally small, with most coefficients near zero across both lexical decision and naming tasks. These findings suggest that while feature norms may capture some aspects of conceptual similarity, they do not consistently or strongly predict priming effects. This shows that feature overlap alone may be insufficient for accurately modeling semantic similarity.
More recently, researchers have begun to explore whether large language models (LLMs) can serve as an alternative for human-based feature norms in predicting semantic priming effects. These models capture word meaning based on text co-occurrence patterns between words extracted form large corpora, such as movie subtitles. The subs2vec model used in this study is an example of such a model [VPT21]. The subs2vec model creates a single static vector for each word based on how frequently that co-occurrence patterns (how frequently words appear near each other) based on the OpenSubtitles dataset, which is a large collection of movie subtitles. This dataset is chosen due to its reflection of everyday, conversational speech in comparison to more formal texts (e.g., Wikipedia), which improves its ability in predicting semantic effects like semantic priming. In this approach, semantic similarity between cue and target is quantified as the cosine similarity between their respective embedding vectors and can be used to predict semantic priming effects in the same was as feature overlap.

2.1.6 Research Questions

Although semantic similarity has been studied extensively using feature norm and large language models (LLMs), few studies have directly compared these methods within the same experimental

setup. While both methods are commonly used to represent semantic similarity, it remains a question whether either can reliably predict semantic priming and how they compare when evaluated on the same set of cue–target word pairs. Previous research has rarely tested both approaches side by side in large and reliable datasets like SPAML. As a result, it remains unclear whether LLM-derived similarity captures similar or distinct aspects of meaning compared to human feature norms, and whether one approach offers unique explanatory value beyond the other.
This thesis aims to address this gap by systematically comparing LLM-based embeddings and human feature norms in predicting semantic priming effects. The central research question is: *How do feature overlap measures from LLMs compare with human feature norms in predicting semantic priming effects?* To answer the research question, this paper examines four sub-questions:

- Do LLM-based similarity measures predict semantic priming effects?
- Do human feature norms predict semantic priming effects?
- Which similarity measure better predicts semantic priming effects?
- Do human feature norms provide unique predictive value beyond LLM similarity?

By comparing these approaches, the thesis aims to better understand how similarity based on human judgments and similarity derived from language models contribute to priming effects. This not only helps clarify how people process meaning in real time, but also offers insights for improving language models that seek to reflect or simulate human-like understanding of word relationships.

2.2 Methodology

This section describes the datasets used in this study and how they were used to examine whether different types of semantic similarity predict semantic priming effects. First, each dataset is introduced and its purpose is explained. Then, the steps taken to process the data and construct the final analysis dataset are discussed. Finally, the statistical analyses used to test the predictive value of the similarity measures are described.

2.2.1 Datasets

This thesis makes use of three datasets to explore whether different measures of semantic similarity can predict semantic priming effects. The primary dataset is SPAML, providing the reaction time data from a large-scale semantic priming experiment. This dataset will be compared to both an LLM-base feature norms and human-based feature norms. All analyses are conducted using the English subset of each dataset.

The SPAML dataset As discussed in Section 2.1.4, SPAML is a large-scale semantic priming dataset [1]. It provides reaction times for a continuous lexical decision task on a wide range of cue–target word pairs. This allows for the calculation of semantic priming by subtracting the mean z-transformed response time to related targets from that of unrelated targets. These priming scores serve as the dependent variable for evaluating whether LLM-based similarity measures and human feature norms can reliably predict priming.

[1] SPAML dataset available on GitHub

The subs2vec dataset The subs2vec dataset[2], introduced in Section 2.1.5, represents the large language model (LLM)-based similarity in this study [VPT21]. Subs2vec computes similarity based on how often words appear in similar contexts across a massive corpus. Specifically, it was trained on the English OpenSubtitles corpus, which contains tens of millions of sentences from movies and TV shows. This corpus is particularly suited for modeling human-like language use, as it reflects a wide range of conversational, everyday speech. For each word, subs2vec creates a high-dimensional vector based on its contextual usage across this corpus. The cosine similarity between these vectors is then used to estimate semantic similarity. In this thesis, cosine similarity scores derived from the subs2vec dataset serve as the LLM-based predictor of semantic priming in the SPAML dataset and will be referred to as LLM-based similarity or predictor.

The feature norms dataset The third dataset, referred to as the feature norms dataset[3] in this thesis, is based on cue-feature norms compiled by Buchanan et al. [BVM19]. These norms were created using a feature listing task, in which participants were shown a cue word (e.g., *zebra*) and asked to list its properties (e.g., *stripes, tail*). For each cue, features were processed both in their original (raw) form and in a translated (root) form, similar to lemmatization, which involves reducing words to their base or dictionary form to combine similar features under a single representation. Semantic similarity was estimated by calculating the cosine similarity between vectors of normalized feature frequencies, using both raw and root forms of the listed features. This thesis uses the cue-feature data, which is only available in English, to investigate whether similarity based on human-generated features can predict semantic priming effects, and how this compares to similarity estimates from language models. In this thesis, cosine similarity scores derived from the feature norms serve as the human-based predictors of semantic priming in the SPAML dataset. Two variants will thus be used: one raw feature norms and root feature norms. These will be referred to as the human-based raw similarity and the human-based root similarity or predictor.

2.2.2 Data Processing

SPAML To ensure accuracy and reliability of the reaction time (RT) data in the semantic priming task, the SPAML dataset pre-processed according to participant-, trial- and item-level exclusion criteria. At the participant level, individuals were excluded if they did not indicate being at least 18 years old, ensuring basic cognitive maturity. Participants who completed fewer than 100 trials were excluded to guarantee sufficient task engagement for reliable estimation. To account for irregular or unstable response patterns, participants were removed if their RT distributions were multimodal, as identified by Silverman's test. Participants were excluded if they showed unnatural behaviour, such as always pressing the same key or mechanically alternating responses. Furthermore, participants with an overall accuracy below 70% were also excluded, as this suggested inattentiveness or poor understanding of the task. Finally, non-native speakers were excluded to reduce variability linked to second-language processing, which can influence both speed and consistency in lexical decisions. At the trial level, responses were excluded if they exceeded the 3000 ms timeout window, were answered incorrectly, were implausibly fast (under 160 ms) or fell outside ± 3 standard deviations from a participant's mean RT. These trial-level exclusions help remove noise caused by lapses in attention, premature responses, or outliers that might interfere with accurately comparing reaction

[2]Subs2vec data available on GitHub
[3]Feature Norms dataset available on GitHub or the website wordnorms.com

times across conditions within the same participant. At the item level, word pairs were excluded if more than 50% of participants responded to them incorrectly, as such high error rates suggest that the item caused confusion, involved ambiguous word meaning or was too difficult for most participants to recognize correctly. An overview of all applied exclusion criteria at each level is summarized in Table 1.

Level	Criterion
Participant-level	1. Participant did not indicate being at least 18 years old. 2. Participant completed fewer than 100 trials. 3. Participant exhibited a multimodal RT distribution (Silverman's test). 4. Participant showed unnatural response patterns: (a) always pressing the same key, or (b) consistently alternating responses. 5. Participant's overall accuracy was below 70%. 6. Participant was a non-native speaker of the test language.
Trial-level	1. Response exceeded the 3000 ms timeout. 2. Trial was answered incorrectly. 3. Response latency was shorter than 160 ms. 4. Response latency exceeded a participant-based threshold of $\|Z\| > 3.0$.
Item-level	1. The item had an error rate above 50% across all participants.

Table 1: Exclusion Criteria at the Participant, Trial, and Item Levels

2.2.3 Final Dataset Construction

After applying the exclusion criteria described in Section 2.2.2, the cleaned SPAML dataset was prepared for analysis by calculating semantic priming effects. Semantic priming effects were calculated by subtracting the aggregated reaction time (RT) for related trials from that of unrelated trials per target word. This resulted in a semantic priming score for each target word, indicating how much faster participants responded when the cue was semantically related. To test whether feature overlap could predict semantic priming, the dataset was extended with two types of predictors:

- **LLM-based similarity:** cosine similarity between cue–target pairs using subs2vec embeddings trained on a large corpus of film subtitles.
- **human-based similarity:** cosine similarity between vectors derived from human-generated semantic feature norms, using both root-translated and raw versions.

2.2.4 Statistical Analysis

The statistical analysis focused on two types of feature overlap predictors as discussed in section 2.2.3: the LLM-based feature overlap predictor and the Human-based feature overlap predictor. The analysis proceeded in two main steps. First, a correlation analysis was conducted to assess if there is a relationship between the a predictor and the dependent variable (the semantic priming effects) and then the strengths of that relationship was measured. Second, linear regression models were

constructed to evaluate how well each feature overlap measure explains variance in semantic priming, both individually and in combination.

Correlation Analysis To evaluate the relationship between the feature overlap predictors and semantic priming, both Pearson and Spearman correlations were computed. Pearson correlation (r) measures the strength and direction of a linear relationship between two continuous variables. In contrast, Spearman's rank correlation (ρ) measures a monotonic relationship between two variables, based on the rank of the data rather than their raw values. Using both methods provides a more robust measure of the relationship between feature overlap and semantic priming, since it cannot be assumed that the relationship is strictly linear. Correlations were calculated separately for each predictor:the LLM-based predictor, and the Human-based (raw and root-translated) predictor.

Regression Analysis To assess whether feature overlap measures significantly explain variance in semantic priming, linear regression models were constructed using combinations of predictors. First, a model was tested that included both human-based predictors (raw and root-translated feature norms). Then, a full model was constructed, including both the human-based and LLM-based predictors. To compare the predictive performance of the regression models, two approaches were used:

- Information Criteria (AIC and BIC), to evaluate model fit, which is how well a model represents the relationship between variables in a dataset, while accounting for complexity.
- Analysis of variance (ANOVA), to test whether adding predictors significantly improved model performance.

First, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were computed for each model. Both criteria assesses models by balancing goodness of fit against model complexity. AIC evaluates how well a model fits the data while discouraging overfitting by penalizing for each additional parameter. BIC is similar, but is stricter than AIC by applying a stronger penalty for models with more parameters, which increases as the size of the dataset grows. As a result, BIC generally favors simpler models, while AIC may favor slightly more complex models if the improvement in fit justifies the added complexity. For both criteria, the model with the lowest AIC or BIC values indicate a better-fitting model.
In addition to these information criteria, analysis of variance (ANOVA) was used to compare models statistically. ANOVA is used to evaluate whether two or more models are significantly different. It compares the variance between models to the variance within models. The null hypothesis assumes there is no difference in model means, thus the models explain the data equally well. The alternative hypothesis states that at least one model performs significantly better than the others. If the variance between models is large relative to the variance within models, the null hypothesis is rejected, suggesting that the more complex model explains significantly more variance and provides a better fit.

2.3 Results

2.3.1 Correlation Analysis

To evaluate the relationship between the different feature overlap measures and semantic priming, both Pearson's r and Spearman's ρ were calculated. The LLM-based predictor showed a Pearson correlation of $r = 0.320$ with semantic priming, with a corresponding p-value of 1.55×10^{-13}. This indicates a moderate positive linear correlation by Cohen's guidelines [Coh88]. This shows that higher similarity scores are thus moderately associated with stronger priming. The human-based raw predictor showed a Pearson correlation of $r = 0.210$ with a p-value of 1.85×10^{-6} and the human-based root predictor showed a Pearson correlation of $r = 0.199$ with a p-value of 6.73×10^{-6}. Both results indicate a small positive linear correlation with semantic priming. This shows that higher similarity scores from the human-based raw and root predictors are weakly associated with semantic priming. While all three predictors show statistically significant correlation with semantic priming, the LLM-based predictor showed the strongest correlation, followed by the human-based raw predictor and human-based root predictor. The results of these calculations are summarized in Table 2

Predictor	**Pearson's r**	**p-value**
LLM-based predictor	0.320	1.55×10^{-13}
Human-based raw predictor	0.210	1.85×10^{-6}
Human-based root-translated predictor	0.199	6.73×10^{-6}

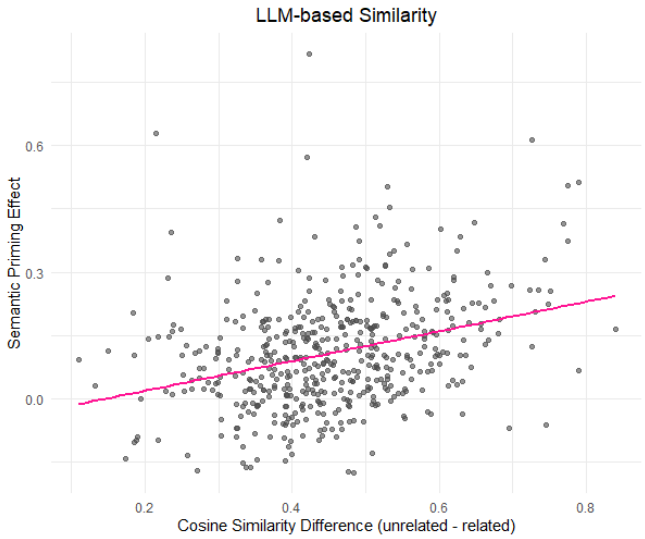
Table 2: Pearson correlation results for the LLM-based similarity and both human-based (raw and root) similarities.

The Spearman's rank correlation confirm this pattern. The results of these calculations are summarized in Table 3. Effect sizes for Spearman's ρ are interpreted using the same benchmarks as Pearson's r, following Cohen's guidelines [Coh88]. The LLM-based predictor showed a Spearman correlation of $\rho = 0.320$, with a p-value of 7.91×10^{-14}, indicating a moderate positive monotonic correlation. This shows that higher similarity scores are thus moderately associated with stronger priming, even when only rank order is considered. The human-based raw predictor showed a Spearman correlation of $\rho = 0.187$ with a p-value of 2.38×10^{-5} and the human-based root predictor showed a Pearson correlation of $r = 0.177$ with a p-value of 6.44×10^{-5}. Both results indicate a small positive monotonic correlation with semantic priming. This shows that higher similarity scores from the human-based raw and root predictors are weakly associated with semantic priming. Comparably to the Pearson results, the LLM-based predictor showed the strongest correlation, followed by the human-based raw predictor and human-based root predictor.

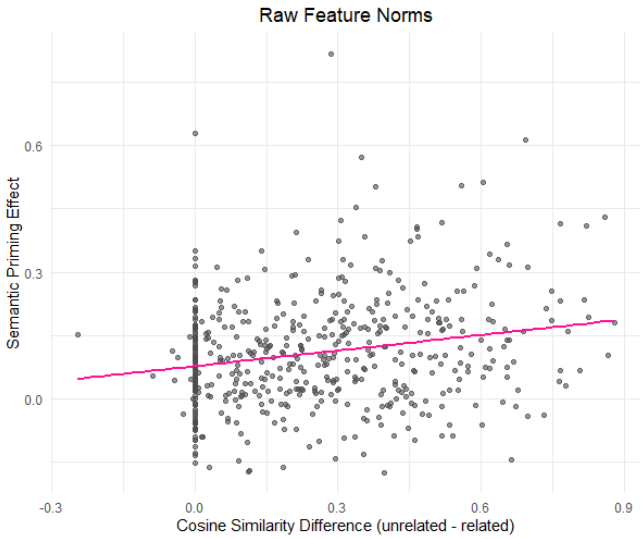
Predictor	**Spearman's ρ**	**p-value**
LLM-based predictor	0.326	7.91×10^{-14}
Human-based raw predictor	0.187	2.38×10^{-5}
Human-based root predictor	0.177	6.44×10^{-5}

Table 3: Spearman correlation results for the LLM-based similarity and both human-based (raw and root) similarities.

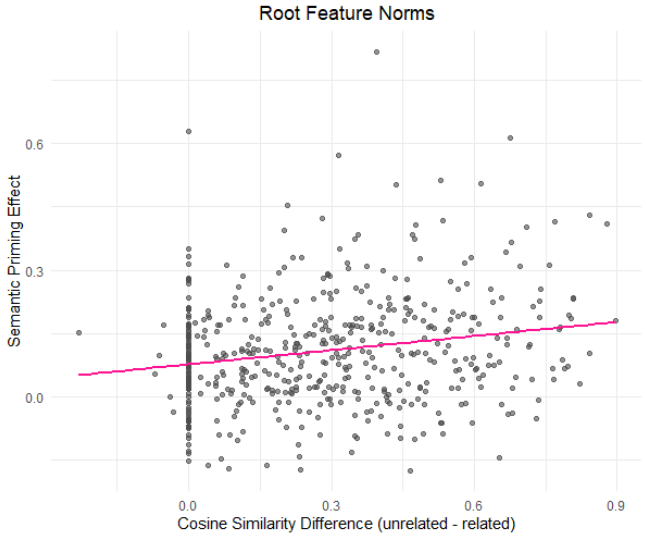
Scatterplots visualizing these relationships are shown in Figure 3. Each subplot includes a linear regression line, illustrating the direction and strength of the observed correlation. Analyzing the scatterplots confirms this pattern once again, with the LLM-based predictor (Figure 3a) showing a denser plot with steeper upward trend, while a difference between the human-based raw and root predictors are hard to distinguish to which is steeper. In contrast, the plot of the human-based raw and root predictors (Figure 3b and 3c) show flatter slow and greater vertical distribution op points, which matches their weaker correlation values. They predictors also show vertical clustering near zero cosine similarity, which occurs because of the many cue-target pairs that share no overlapping features in the feature norms dataset. Despite having zero feature overlap, these pairs still have a semantic priming effect, which is reflected by the vertical difference in height across the zero line. Taken together, the correlation results show that all three predictors are positively correlated to semantic priming. However, the LLM-based predictor showed the strongest, while still moderate, correlation across both the Pearson and the Spearman coefficients.



(a) LLM-based similarity scores



(b) human-based (raw) similarity scores



(c) human-based (root) similarity scores

Figure 3: Scatterplots showing the relationship between cosine similarity difference (unrelated minus related) and the semantic priming effect (z-score difference) across predictors. A linear regression line is shown in pink.

2.3.2 Regression Models

To further evaluate the contribution of LLM-based and human-based feature overlap predictors to semantic priming, a set of linear regression models was constructed and compared using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values for each model, which are presented in Table 4. Among the single-predictor (i.e., LLM-based, human-based raw and human-based root) models, the LLM-based model achieved the lowest AIC (−666.78) and BIC (−654.10) values, suggesting it offered the best individual model fit to the data. In contrast, the human-based raw model (AIC =−634.86 and BIC = −622.18) and human-based root model (AIC = −632.37 and BIC = −619.69) performed substantially worse on both metrics. Adding the human-based raw predictor to the LLM-based model (LLM + raw model) resulted in a lower AIC value (−672.11), indicating a better fit despite its increased complexity, while the BIC increased slightly (−653.74), which reflect its stricter penalty for complexity in models. When the human-based root predictor was added to the LLM-based model (LLM + root model), both the AIC (−632.86) and BIC (−615.69) worsened in comparison to the LLM-based model. This suggests that the human-based root predictor did not significantly improve the model and may introduce unnecessary complexity. The final model that included all three predictors (i.e., LLM-based, Human-based raw and human-based root) achieved a slightly worse AIC (−670.21) and BIC (−649.08) compared to the LLM + raw model, implying that the root predictor, again, did not significantly improve the model and may introduce unnecessary complexity. Together, these models comparisons indicate that the LLM-based predictor provides the best individual fit and adding the human-based raw predictor improves model performance, reflected by the AIC. The human-based root predictor contributed little explanatory value.

Model	**AIC**	**BIC**
LLM-based model	-666.78	-654.10
Human-based raw model	-634.86	-622.18
Human-based root model	-632.37	-619.69
LLM + raw model	-672.11	-653.74
LLM + root model	-632.86	-615.69
LLM + raw + root model	-670.21	-649.08

Table 4: Model comparison using AIC and BIC for LLM-based and human-based predictors, with lower values indicating a better model fit.

In addition to the results found by the AIC and BIC comparisons, ANOVA tests were conducted to assess whether the differences in model fit were statistically significant. The results are summarized in Table 5. Adding the human-based raw predictor to the LLM-based model (LLM + raw model), significantly improved model fit, with an F-statistic of $F(1, 503) = 7.34$ and $p = 0.007$. This indicates that the human-based raw similarity significantly explains additional variance in semantic priming beyond what is already captured by the LLM-based predictor alone. Adding the human-based root predictor to the LLM-based model (LLM + root model) also showed significant improvement to the model fit, with $F(1, 503) = 5.86$ and $p = 0.016$. Although both human-based predictors independently contributed significantly to the LLM-based model, adding the human-based raw predictor to the LLM-based model achieved a higher F-value, which indicates that it provided more explanatory value than adding the human-based root predictor to the LLM-based model. When

both human-based predictors were added to the LLM-based model, the improvement remained statistically significant, with $F(2, 502) = 3.71$ and $p = 0.025$, but the effect was smaller than the other two models. Lastly, when the root predictor was added to the human-based raw model, there was nog significant improvement in model fit, with $F(2, 502) = 0.0015$ and $p = 0.969$, which confirms that the root predictor adds no significant value beyond the human-based raw predictor.

Comparison	**F-statistic (df)**	**p-value**
LLM-based model vs. LLM + raw model	F(1, 503) = 7.34	0.007
LLM-based model vs. LLM + root model	F(1, 503) = 5.86	0.016
LLM-based model vs. LLM + raw + root model	F(2, 502) = 3.71	0.025
Human-based raw model vs. raw + root model	F(1, 503) = 0.0015	0.969

Table 5: ANOVA comparisons between linear regression models evaluating whether the inclusion of additional predictors (human-based raw or root similarity) significantly improves model fit beyond the LLM-based predictor and additionally or raw feature overlap alone.

2.4 Discussion

This study focused on comparing LLM-based feature overlap to human feature overlap in an effort to determine which measure is a better predictor of semantic priming. The findings in Section 2.3 show that both predictor, LLM-based feature overlap, human feature overlap (both raw and root form), significantly correlate with semantic priming effects (Table 2 and 3). Among these, LLM-based similarity exhibited the strongest correlation with semantic priming, followed by human-based similarity in raw form and human-based similarity in root form. These results suggest that both distributional and human feature-based representations capture meaningful aspects of semantic similarity, though to varying degrees. Regression analyses further support this interpretation. When comparing models that included each predictor individually, the LLM-based similarity showed the lowest AIC and BIC values (Table 4), indicating that LMM-based similarity provided the best overall fit to the data. Adding the raw human-based predictor to the LLM-based predictor, further reduced AIC, suggesting an improved explanatory value, although BIC slightly increased due to its stronger penalty for added complexity. Because the decrease in AIC outweighs the increase in BIC, the combined model was still preferred overall. This interpretation is further supported by the ANOVA results (Table 5), which confirmed that including raw feature overlap significantly improved the models ability to explain variance in semantic priming. These results suggest that while LLM-based similarity capture the strongest and most consistent patterns in the data, human feature norms still contribute a unique explanatory value, which is not captured in LLM-based similarities.

2.5 Conclusion

To answer the main research question, 'How do feature overlap measures from LLMs compare with human feature norms in predicting semantic priming?', this study examines four sub-questions.
priming effects?
First, we found that LLM-based feature overlap measures significantly predicted semantic priming effects, showing the strongest individual correlation with semantic priming. Second, human feature

norms, both in raw and root-translated form, also significantly predicted semantic priming, although the raw form perform slightly better. Third, when directly compared, the LLM-based feature overlap was the better overall predictor. Finally, raw and root-translated feature norms improved the predictive value of the LLM-based model when added separately. However, once one one human-based predictor was included, the other no longer explained additional variance, suggesting that they capture largely overlapping information. These findings indicate that LLMs are effective at modeling semantic similarity and predicting semantic priming, but that human feature norms still contribute unique conceptual information. Combining both approaches offers a more comprehensive understanding of semantic similarity.

3 Study 2: Extending Semantic Priming Prediction with Association Strength

3.1 Introduction

The first study in this thesis compared two types of semantic similarity: feature-based similarity and LLM-based similarity. While both were significantly correlated with semantic priming, the study showed that the LLM-based similarity performed better in predicting semantic priming. However, it remains possible that other forms of human-derived similarity might better predict semantic priming effects. This study introduces a third type of semantic similarity measure: association strength. Building further on the theories discussed in Study 1, this section will first introduce associations, how associative similarity is measured and will introduce a new dataset.

3.1.1 Associations and Their Role in Predicting Semantic Priming

As introduced in Section 1, associations refer to the connection between words in the mind. They have been widely studied through free association tasks, where an individual is presented with a word (e.g., *stork*) and has to respond with the first word that comes to mind (e.g., *baby*). Despite its simplicity, the task is considered to be powerful, since generating a word associate for a cue is easy and since it proves difficult to not respond with the first thing that comes to mind [DDNP+19]. Importantly, word associations reveal conceptual links in the mental lexicon that go beyond co-occurrence patterns in the language. Since the task does not depend on communication in natural language, it can capture mental representations that are not represented in lexical patterns used by models. Word association tasks provide association strength values, which is the probability of responding with a certain word, given a specific cue. We consider two types of association strength, namely forward association strength and backward association strength [Hut03]. Forward association refers to how likely a cue word (e.g., *stork*) triggers a particular target word (e.g., *baby*) in a free association task. Backward association measures the reverse, how likely, in this case, *baby* triggers the word *stork*. These associations are often asymmetrical, which means that a strong forward association does not necessarily imply a strong backward association. Spreading activation theories predict stronger priming in the forward direction, since activation flows from the cue words to associated concepts. To study associative similarity, researches have relied on large-scale free association norms. An example of such a dataset is the Small World of Words (SWOW) project [DDNP+19]. SWOW is a large-scale, crowd-sourced free association (task) project that has collected

word association responses from thousands of participants across multiple languages. SWOW uses a continued-response procedure, allowing participants to provide up to three associations per cue. This does not only increase the likelihood of capturing weak associations, but also helps enabling the construction of denser semantic networks. Together, the scale and methodological design of SWOW make it a valuable resource for capturing both strong and weak associative links.

3.1.2 Research Question

While there have been prior studies systematically comparing the role of associative strength to feature norms in semantic priming [Hut03], few have directly compared associative similarity to both feature norms and LLM-based similarity measures within a unified analysis. Moreover, such comparisons have typically relied on small-scale or less reliable datasets, limiting the generalizability of their conclusions. In contrast, the SPAML dataset offers a large and highly reliable resource for examining priming effects across a wide range of stimuli, using a consistent experimental approach. Building on Study 1, which showed that LLM-based similarity outperformed feature norms in predicting priming effects, this second study aims to evaluate whether associative similarity adds predictive value beyond these two approaches. By incorporating associative strength into the same analytic framework, this study provides an empirical comparison of all three types of semantic similarity within the SPAML dataset. The central research question is: *Does associative similarity offer unique predictive value for semantic priming effects beyond what is captured by LLM-based similarity and human feature norms?* To answer this, the following sub-questions are addressed:

- Does associative similarity significantly predict semantic priming effects?
- How does associative similarity compare to LLM-based and feature norm-based similarity in terms of predictive strength?
- Does associative similarity provide unique explanatory power when combined with LLM-based and feature norm predictors?

3.2 Methodology

Study 2 builds directly on the methodology established in Study 1. Unless specified otherwise, all procedures concerning data processing, exclusion criteria, dataset construction and statistical analysis remain unchanged. The main methodological addition in this study is the inclusion of a third predictor: associative similarity, derived from the SWOW dataset, described in Section 3.1.1. This addition allows for further evaluation of human association strength, as a potential explanatory measure in predicting semantic priming effects.

3.2.1 Datasets

This study uses the same datasets as Study 1. The SPAML dataset was used for priming scores, the subs2vec dataset was used to to simulate the LLM-based feature overlap and the feature norms dataset was used human-based feature overlap. In an associative similarity measure was included, based on the Small World of Words dataset, introduced in Section 3.1.1

The Small World of Words dataset The Small World of Words (SWOW) dataset[4] contains word association data collected through a large crowd-sourced experiment by De Deyne et al. [DDNP+19]. Participants were shown a cue word and asked to write down the first three words that came to mind. The procure allowed for cues with multiple responses, resulting in the ability to capture both strong and weak associativity.

The human feature overlap dataset Based on the findings from Study 1, Study 2 will only include the raw feature norms as the human-based predictor. This decision is to reduce complexity on the basis that root-translated feature overlap not explaining addition variance beyond raw feature overlap.

Additionally, to avoid confusion, in this study the LLM-based feature overlap is called LLM-based similarity or predictor, the human feature overlap is called feature overlap-based similarity or predictor and the association strength is called association-based similarity or predictor.

3.2.2 Final Dataset Construction

Following the same filtering procedure as in Study 1, semantic priming effects were calculated by subtracting z-transformed reaction times for both related and unrelated target words. The final dataset for study 2 was extended with associative similarity from SWOW, in addition to the existing LLM-based and human feature overlap. As a result of this stricter requirement, where both the cue and target words needed to be present across all three predictor datasets, the total number of target words with a corresponding related and unrelated cue was reduced from 506 in Study 1 to 246 in Study 2. Comparable to the feature norms dataset in Study 1, an exception was made for the SWOW data: if the unrelated cue was not available, but the related cue and target were, the associative similarity score was set to zero. This allowed for more word pairs to be retained, based on the assumption that missing data for the unrelated cue reflects a lack of associative connection with the target word. Without this exception, only 157 out of 246 target words (63.8 %) would have had both a related and unrelated cue with valid association strength values. This dataset formed the basis for all correlation, regression and model comparison analyses in Study 2.

3.2.3 Statistical Analysis

The statistical analyses in Study 2 followed the same procedures as in Study 1. Since the dataset is smaller, all analyses were re-run using the reduced set of cue-target pairs. The main addition in Study 2 is the inclusion of the associative similarity predictor, which results in a direct comparison between associations, human feature overlap and LLM-based feature overlap in predicting semantic priming.

3.3 Results

3.3.1 Correlation Analysis

To evaluate the relationship between the different predictors and semantic priming, both Pearson's r and Spearman's ρ were calculated. The LLM-based predictor showed a Pearson correlation of

[4]SWOW dataset available at smallworldofwords.org

$r = 0.339$ with semantic priming, with a corresponding p-value of 4.79×10^{-8}. This indicates a moderate positive linear correlation by Cohen's guidelines [Coh88]. This shows that higher LLM-based similarity scores are moderately associated with stronger priming effects. The feature overlap predictor showed a Pearson correlation of $r = 0.213$ with semantic priming, with a corresponding p-value of 7.77×10^{-4}, indicating a small positive linear correlation. In contrast, the associations predictor showed a Pearson correlation of $r = 0.052$ with semantic priming, with a corresponding p-value of 0.416, suggesting no significant linear correlation with semantic priming. The results are summarized in Table 6.

Predictor	**Pearson's r**	**p-value**
LLM-based predictor	0.339	4.79×10^{-8}
Human predictor (raw)	0.213	7.77×10^{-4}
Associations predictor	0.052	0.416

Table 6: Pearson correlation results for the LLM-based features norms, raw human feature norms and association similarity.

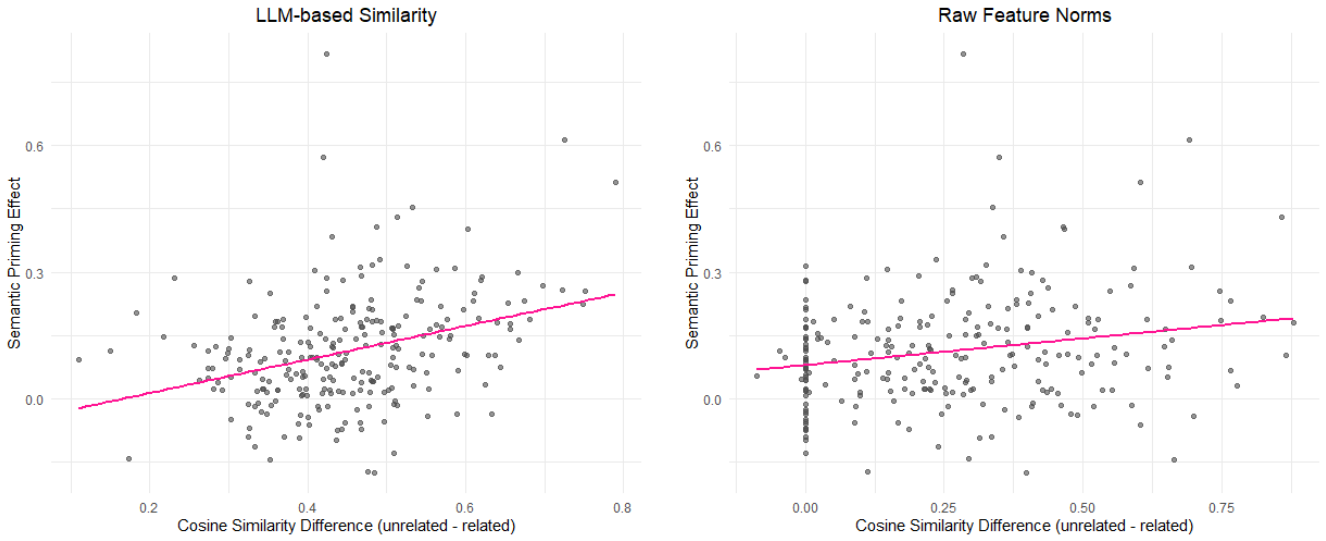
The spearman correlation confirmed this pattern. The results are summarized in Table 7. Effect sizes for Spearman's ρ are interpreted using the same benchmarks as Pearson's ρ, following Cohen's guidelines [Coh88]. The LLM-based predictor showed a Spearman correlation of $\rho = 0.349$, with a p-value of 2.35×10^{-8}, indicating a moderate positive monotonic correlation. This shows that higher LLM-based similarity scores are moderately associated with stronger priming, even when considering only the rank order of values. The feature overlap-based predictor showed a Spearman correlation of $\rho = 0.191$ with a p-value of 2.66×10^{-3}, indicating a small positive monotonic correlation. The association-based predictor again showed no significant correlation with $\rho = 0.104$ and a p-value of 0.104. These results suggest that while both LLM-based and feature overlap-based similarity measures significantly correlate to semantic priming, associative similarity is not.

Predictor	**Spearman's ρ**	**p-value**
LLM-based predictor	0.349	2.35×10^{-8}
feature overlap-based predictor	0.191	2.66×10^{-3}
Association-based predictor	0.104	0.104

Table 7: Spearman correlation results or the LLM-based features norms, raw human feature norms and association similarity.

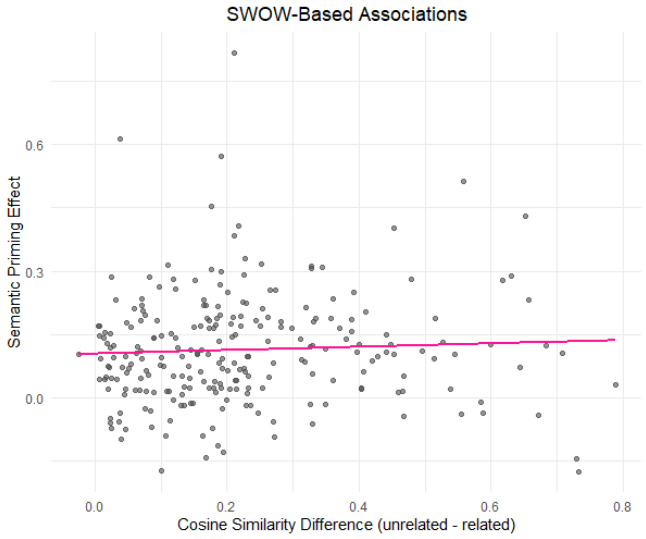
Scatterplots visualizing these relationships are shown in Figure 4. Each subplot includes a linear regression line, illustrating the direction and strength of the observed correlation. For the LLM-based predictor (Figure 4a), a visible cluster of points follows the direction of the regression line, supporting the observed moderate positive correlation. The scatterplot for the feature overlap-based (Figure 4b) also shows, although less pronounced, a loose clustering along the regression line. The figure also show vertical clustering near zero cosine similarity, which occurs because of the many cue-target pairs that share no overlapping features in the feature norms dataset. Despite having zero feature overlap, these pairs still have a semantic priming effect, which is reflected by the vertical difference in height across the zero line. In contrast, the associations-based predictor (Figure 4c) shows a distributed pattern. Most data points are concentrated at the lower end of

the cosine similarity scale and the upward trend is more subtle, which aligns with the results that association-based similarity does not significantly predict semantic priming.



(a) LLM-based similarity scores

(b) Feature-overlap-based similarity scores



(c) Association-based similarity scores

Figure 4: Scatterplots showing the relationship between cosine similarity difference (unrelated minus related) and the semantic priming effect (z-score difference) across predictors. A linear regression line is shown in pink.

3.3.2 Regression Models

To further evaluate the contributions of each predictor to semantic priming, a set of linear regression models was constructed and compared using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values. The AIC and BIC values are presented in Table 8 for each model. Among the single-predictor models, the LLM-based model achieved the lowest AIC (-324.86) and BIC (-314.34) values, suggesting it offered the best individual model fit to the data. The feature overlap-based model performed notably worse (AIC = -306.16 and BIC = -295.65), as association-based model (AIC = -295.42 and BIC = -284.91), indicating that these predictors explained less variance in semantic priming when used on their own. When the feature overlap-based predictor was added to the LLM-based model (LLM + feature overlap model), the AIC (-325.29) improved slightly, suggesting a marginal in the explanation of variance of semantic priming. However, the BIC (-311.27) showed an increase in value, reflecting the strict penalty for model complexity. Adding the association-based predictor to the LLM-based model (LLM + association model) resulted in an increase in both AIC and BIC (AIC = -323.03 and BIC = -309.01), suggesting that association similarity did not meaningfully improve the model and introduced unnecessary complexity. Similarly, combining the association-based predictor to the feature overlap model (feature overlap + association model) also resulted in an increase in both AIC and BIC (AIC = -305.74 and BIC = -291.72), suggesting that association similarity, again, did not meaningfully improve the model and introduced unnecessary complexity. The full model, which included all three predictors, achieved the lowest AIC overall (-325.35), but the BIC was higher than both the LLM-based model and the LLM + feature overlap model, suggesting that the added complexity may not be justified. Together, these model comparisons indicate that the LLM-based predictor remains the strongest individual predictor of semantic priming in this dataset. While adding the feature overlap-based predictor slightly improves the AIC, the BIC suggests that the added complexity may not be justified. The association-based predictor, whether added alone or in combination with other predictors, does not improve model fit.

Model	**AIC**	**BIC**
LLM-based model	-324.86	-314.34
feature overlap-based model	-306.16	-295.65
Association-based model	-295.42	-284.91
LLM + feature overlap model	-325.29	-311.27
LLM + association model	-323.03	-309.01
feature overlap + association model	-305.74	-291.72
LLM + feature overlap + association model	-325.35	-307.82

Table 8: Model comparison using AIC for LLM-based similarity, feature overlap and associative similarity predictors, with lower values indicating better model fit.

In addition to the AIC and BIC analysis, ANOVA tests were conducted to assess whether the differences in model fit were statistically significant. The results summarized Table 9 and show that none of the comparisons achieved statistical significance, which indicated that adding addition predictors did not significantly improve the models. Specifically, adding feature overlap to the LLM-based model did not result in a significant difference ($F = 2.42$, $p = 0.121$). Likewise, adding the association predictor to the LLM-based model ($F = 0.17$, $p = 0.680$), to the feature overlap-based

model ($F = 1.56$, $p = 0.213$) or to the LLM + feature overlap model ($F = 2.03$, $p = 0.156$) did not lead to a significant difference. Lastly, adding both feature overlap and the association predictor to the LLM-based model did also not lead to a significant difference ($F = 2.23$, $p = 0.110$). Together, these ANOVA results reinforce the conclusions drawn from the AIC and BIC comparisons: the LLM-based similarity measure is the strongest individual predictor of semantic priming in this study. Neither feature overlap nor associative similarity significantly improved model performance when added individually or in combination.

Model comparison	**F-statistic**	**p-value**
LLM vs. LLM + feature overlap	F(1, 243) = 2.42	0.121
LLM vs. LLM + association	F(1, 243) = 0.17	0.680
feature overlap vs. feature overlap + association	F(1, 243) = 1.56	0.213
LLM + feature overlap vs. LLM + feature overlap + association	F(1, 242) = 2.03	0.156
LLM vs. LLM + raw + association	F(2, 242) = 2.23	0.110

Table 9: ANOVA comparisons between linear regression models evaluating whether the inclusion of associative similarity significantly improves model fit beyond LLM-based and/or human-based feature overlap predictors.

3.4 Discussion

This study extended the comparison between LLM-based and human-based feature overlap by introducing associative similarity as a third predictor of semantic priming. The findings in Section 3.3 show that both the LLM-based feature overlap and the human feature overlap in its raw form significantly correlate with semantic priming (Table 6 and 7). Among these, the LLM-based feature overlap again showed the strongest correlation with semantic priming, followed by human feature overlap. Associative similarity, as derived from the SWOW dataset, did not show a statistically significant correlation with semantic priming. These results suggest that while both the LLM-based similarity and feature overlap significantly capture meaningful aspect of semantic similarity, thus being able to significantly predict semantic priming, associative similarity does not appear to contribute in significantly capturing aspects of semantic similarity. Regression analyses further support this, since among the individual models, the LLM-based predictor achieved the lowest AIC and BIC values 8, indicating the best model fit. When the feature overlap-based predictor was added to the LLM-based model, the AIC decreased slightly, suggesting a modest improvement in explanatory power. However, the BIC increased due to its stricter penalty for added complexity. In contrast, when adding associative similarity to either the LLM-based model, the feature overlap model or the combined (LLM + feature overlap) model resulted in an increase in both AIC and BIC. This indicates that associative similarity reduces model performance. The full model, which includes all three predictors, achieved the lowest AIC value overall, but has a higher BIC, which suggests that the added complexity was not justified. This has been supported by the ANOVA results that showed that none of the model comparisons achieved statistical significance (Table 9). This suggests that while LLM-based similarity consistently explains a significant portion of the variance in semantic priming, neither human feature overlap nor associative similarity provides a statistically significant contribution in this smaller dataset. Together, these results show that LLM-based

similarity is the most robust predictor of semantic priming. While human feature overlap may offer some additional explanatory value, the contribution is small. Associative similarity, despite its longstanding theoretical relevance, did not provide any addition explanatory value when evaluated alongside both human-based and LLM-based feature overlap.

3.5 Conclusion

To answer the main research question, *Does associative similarity offer unique predictive value for semantic priming effects beyond what is captured by LLM-based similarity and human feature norms?*, this study addressed three sub-questions. First, we found that associative similarity, as measured by the SWOW-dataset, did not significantly predict semantic priming effects in the SPAML dataset. In contrast, both LLM-based similarity and human-based similarity did show significant correlations with priming. Second, when directly compared, LLM-based similarity remained the strongest individual predictor of semantic priming, followed by human-based similarity. Third, adding associative similarity to models that already included LLM-based or human-based predictors did not provide any unique explanatory value. These findings indicate that associative similarity does not predict semantic priming nor adds predictive value to what is already captured by LLM- or feature-based similarities, while LLM-based models continue to perform the best.

4 General Discussion

4.1 Reflection on Results

The aim of this thesis was to compare different measures of semantic similarity in predicting semantic priming. Across two studies, three types of similarity were tested: LLM-based feature overlap, human-based feature overlap and associative similarity. The results across both studies consistently showed that LLM-based feature overlap, modeled by subs2vec, was the strongest predictor of semantic priming. This result aligns with recent research showing that LLMs can accurately perform semantic tasks, including predicting semantic priming. [CBA+23]. In both Study 1 and Study 2, LLM-based similarity showed the highest correlational with semantic priming and provided the best model fit according to the AIC and BIC comparisons. This suggests that LLM-based similarity model best explained the variance in semantic priming better in comparison to both human-based feature overlap predictor and associative similarity predictor. Human-based feature overlap also significantly predicted semantic priming in both studies. In Study 1, when human-based similarities were added to the LLM-based model it also significantly improved the model and added explanatory value, which is confirmed by both the AIC/BIC and ANOVA results. These results suggest that while LLM-based similarity capture the strongest and most consistent patterns in the data, human feature norms still contribute a unique explanatory value, which is not captured in LLM-based similarities. However, Study 2 did not show the same result. Study 2 had a smaller dataset (amount of cue-target pairs) due to the inclusion of associative similarity, which may have lead to the loss of explanatory value, since human feature overlap did no longer improve the LLM-based model significantly. Furthermore, root-translated feature norms never added value beyond the raw form, which supports Buchanan et al.'s [BVM19] observation that morphological details (e.g., affixes) often carry semantically relevant distinctions. Associative similarity, derived

from forward association strength in the SWOW dataset [DDNP+19], did not significantly correlate with semantic priming and did not improve model performance when added to either LLM-based or (human) feature overlap predictor in Study 2. This is in contrast with the expectations of De Deyne et. al.[DDNPS12], which emphasizes the importance of word associations. The current findings of this thesis suggest that forward association strength, when implemented through cosine similarity, is not sufficient to capture semantic relationships between words and predict semantic priming. While the SWOW dataset remains valuable for mapping mental representations, it did not provide predictive value for this task. Together, the results support the conclusion that LLM-based models offer strong correlative and predictive value for semantic priming. Human feature norms can still contribute useful conceptual information, but their added value may depend on dataset size and lexical coverage. In contrast, associative similarity does not appear to explain variance in semantic priming.

4.2 Limitations

Several limitations should be considered. First, while the LLM-based feature overlap and the SPAML dataset come from the same source [BCC+21], the human feature norms [BVM19] and associative similarity[DDNP+19] come from separate sources. As a result, only cue-target pairs that were present across all predictors (included in the corresponding study) could be included in the analysis[5]. This may have affected the predictive strength of the human-based feature overlap in Study 2, since the exploratory value feature-overlap added to the LLM-based model disappeared when the dataset became smaller. Second, while the datasets used in this thesis are sufficiently large enough to allow for a meaningful analysis, they are still limited in size due to a lack of coverage between predictors. To maximize data retention, target words with corresponding related and unrelated cue words, with a missing unrelated cue in either the feature norms or SWOW dataset were retained by assigning a similarity score of zero. This increased the number of analyzable cue-target pairs. For example, in study 1, the analyzable target words increased from 327 to 506, which is an increase of 54.74% and in Study 2 the analyzable target words increased from 157 to 246, which is a 56.69% increase. However, this also introduced the assumption that missing values reflect no similarity, which may not always be accurate. Some of these unrelated cues may have had a weak but nonzero similarity if they were available. This could affect the reliability of the results and underestimate the predictor's true explanatory strength. Third, all similarity scores in this thesis were based on cosine similarity, which only captures directional overlap between two vectors. However, research by De Deyne et al. [DDNPS12] suggests that more global similarity measures, such as random walk similarity, can better reflect human intuition, particularly for weakly related word pairs, due to them taking more data into account. Thus, the lack of statistical significance for associative similarity in Study 2 may reflect a limitation of the cosine metric rather than an absence of predictive value.

4.3 Future Research

For further research, while the LLM-based predictors showed the strongest performance overall, it would be of additional value to test whether it performs equally well across different semantic

[5]With an exception of certain unrelated cue words, which is explained in Section 2.2.3 and 3.2.2.

categories. For instance, LLMs may be particularly effective a predicting nouns like animals (e.g., *cat-dog*), but less effective for abstract words (e.g., *truth - belief*) Future research could group cue-target pairs by semantic category and analyze whether predictive performance of LLMs differ across domains, helping to determine the how well LLMs predict across domains. Next, newer or alternative LLMs could be tested. This thesis uses subs2vec, which is trained on movie subtitles. While this offers advantages over more formal corpora like Wikipedia [VPT21], future studies could explore whether modern models like GPT-4, that use contextual embeddings, could be evaluated to see whether they offer improvements over static word vectors. Additionally, future studies could explore whether models trained on domain-specific texts (e.g., medical) show an increase in predictive performance in predicting semantic priming on a dataset in the same field. Lastly, future research could address several limitations. First, the current study relied on cosine similarity. Network-based similarity measure (e.g., path length or random walk similarity) could capture indirect or less obvious semantic relationships that cosine similarity might miss [DDNPS12]. In addition, the worse predictive performance of human feature overlap and associative strength may have been underestimated due to the lack of vocabulary coverage between the datasets. Expanding these datasets to include more words and collecting more features or associations per word could improve their predictive strength.

5 General Conclusion

In conclusion, this thesis shows that both LLM-based similarity and feature norms are significantly related to semantic priming effects, with LLMs offering the strongest predictive power. Feature norms contribute modest additional value, suggesting they still capture unique conceptual information. Associative similarity, in contrast, did not improve predictions under current conditions. Together, these results support the idea that language models effectively capture patterns of meaning relevant for real-time word recognition, while highlighting the continued relevance of human-generated semantic knowledge for enriching our understanding of lexical processing.

6 Use of Generative AI

Generative AI was used solely as a writing aid throughout this thesis. Specifically, ChatGPT[6] was used to assist with LaTeX formatting, grammar correction, sentence rephrasing for clarity and finding more suitable synonyms. It served as a tool to support clarity and consistency in writing, but was not used to generate any original content, conduct analyses, interpret results or form arguments. All ideas, interpretations, and conclusions are my own.

[6]The chats can be found here: 1, 2, 3, 4.

References

[BCC+21] Erin Michelle Buchanan, Kelly Cuccolo, Nicholas Alvaro Coles, Tom Heyman, Aishwarya Iyer, Neil Anthony Lewis, Kim Olivia Peters, Niels van Berkel, Anna Elisabeth van't Veer, Jack Edward Taylor, et al. Measuring the semantic priming effect across many languages. 2021.

[BVM19] Erin M Buchanan, Kathrene D Valentine, and Nicholas P Maxwell. English semantic feature production norms: An extended database of 4436 concepts. *Behavior Research Methods*, 51:1849–1863, 2019.

[CBA+23] Giovanni Cassani, Federico Bianchi, Giuseppe Attanasio, Marco Marelli, and Fritz Guenther. Meaning modulations and stability in large language models: An analysis of bert embeddings for psycholinguistic research, Oct 2023.

[CL75] Allan M. Collins and Elizabeth F. Loftus. A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407–428, 1975.

[Coh88] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd ed. edition, 1988.

[CQ69] Allan M. Collins and M. Ross Quillian. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8(2):240–247, 1969.

[DDNP+19] Simon De Deyne, Danielle J Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. The "small world of words" english word association norms for over 12,000 cue words. *Behavior research methods*, 51:987–1006, 2019.

[DDNPS12] Simon De Deyne, Daniel Navarro, Amy Prefors, and Gert Storms. Strong structure in weak semantic similarity: A graph based account. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34, 2012.

[HBN+13] Keith A Hutchison, David A Balota, James H Neely, Michael J Cortese, Emily R Cohen-Shikora, Chi-Shing Tse, Melvin J Yap, Jesse J Bengson, Dale Niemeyer, and Erin Buchanan. The semantic priming project. *Behavior research methods*, 45:1099–1114, 2013.

[Hut03] Keith A Hutchison. Is semantic priming due to association strength or feature overlap? a microanalytic review. *Psychonomic bulletin & review*, 10:785–813, 2003.

[MCSM05] Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559, 2005.

[MKB17] Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92:57–78, 2017.

[MS71] David E Meyer and Roger W Schvaneveldt. Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of experimental psychology*, 90(2):227, 1971.

[Nee77] James H Neely. Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of experimental psychology. General*, 106(3):226–254, 1977.

[SW08] Brigitte Stemmer and Harry A. Whitaker. *Handbook of the Neuroscience of Language*. Academic Press, 2008.

[VPT21] Jeroen Van Paridon and Bill Thompson. subs2vec: Word embeddings from subtitles in 55 languages. *Behavior research methods*, 53(2):629–655, 2021.