Modeling semantic association in self-paced reading with language model embeddings

Sara Møller Østergaard*, **Kenneth Enevoldsen**†, **Afra Alishahi***, **and Bruno Nicenboim***
*Department of Computational Cognitive Science, Tilburg University
†Center for Humanities Computing, Aarhus University
s.m.ostergaard@tilburguniversity.edu

Abstract

Semantic association between a word and its context has been identified as an important component of reading comprehension, even when word predictability is accounted for. Recent research has highlighted the potential of language model (LM) embeddings to quantify semantic association. Yet, embedding-based semantic association have been operationalized in a myriad of ways. In this study, we use embeddings from LMs to estimate semantic association on a corpus of joint electroencephalography (EEG) and self-paced reading of natural, Dutch texts. Semantic association is calculated in ten different implementations that vary the embedding model and context lengths. The effects of semantic association across the different implementations on the N400 and self-paced reading times are examined using Bayesian hierarchical models and Bayes factor. The results show that the choice of embedding model can alter the estimated effect of semantic association on both the N400 and self-paced reading times. Furthermore, the results demonstrate a promising potential of sentence embeddings for capturing semantic association, as only implementations relying on sentence embeddings indicate reliable results of semantic association beyond word predictability on both neural and behavioral measures. Together, these findings highlight the importance of methodological choices in quantifying semantic association.

Keywords: semantic association, self-paced reading (SPR), electroencephalography (EEG), N400, sentence processing

1. Introduction

Humans process words in the context in which they are presented. How predictable a word is given its preceding context largely impacts the processing difficulty of the word (Kutas and Federmeier, 2011; Ehrlich and Rayner, 1981; Wong et al., 2024). For example, in the sentence pair, "By the end of the day, the hiker's feet were extremely cold and wet. It was the last time he would ever buy a cheap pair of *boot/jeans*.", the final word "boots" is highly predictable based on the preceding context and is therefore processed more easily than the alternative ending "jeans", which is comparatively unpredictable in this context (example from Federmeier and Kutas, 1999).

Predictability of a word, or its probability given a context, has been estimated using a range of probabilistic models, including probabilistic grammars (Hale, 2001) and, more recently, next token probabilities from language models (LMs) (Broderick et al., 2018; Ettinger et al., 2016; Xu et al., 2024; Michaelov et al., 2024; Frank, 2017; Michaelov and Bergen, 2024). Additionally, word predictability has been estimated using the cloze task[1] (Luke and Christianson, 2018; Dambacher et al., 2006; Bulkes et al., 2020).

Word predictability has been able to explain important aspects of processing difficulty, however, it doesn't provide a full account. In addition to predictability, semantic association presents another factor that modulates reading comprehension (Kutas and Federmeier, 2011; Brouwer et al., 2012). Semantic association refers to the degree of semantic relatedness between a target word and the context in which it is presented. While this measure is related to the predictability of the word, it has distinct properties. Using the example context from above, "By the end of the day, the hiker's feet were extremely cold and wet. It was the last time he would ever buy a cheap pair of *sandals*.", the word "sandals" is unpredictable in the context, however, it is semantically associated with the context (which mentions feet). Federmeier and Kutas (1999) show that this distinction results in different processing of these target words.

Semantic illusion has been used to study the effects of semantic association beyond word predictability. Semantic illusions refer to a phenomenon where unpredictable (or incorrect) words are temporally unnoticed because the words are semantically associated with the context. The sentence "For breakfast the *eggs* would only eat toast and jam.", illustrates this effect, where the word "eat"

[1]The cloze task is a language comprehension task in which one or more words are removed from a text and must be filled in by the participant based on contextual cues.

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 60–73
16 May 2026.

fails to elicit the expected neural response to an unpredictable word (Kuperberg et al., 2003). Studies on semantic illusion report that words semantically associated with their context are processed differently (as shown with electroencephalography; EEG) compared to words that lack such associations (Kuperberg et al., 2003; Nieuwland and Van Berkum, 2005; Stone and Rabovsky, 2025; Aurnhammer et al., 2023). Relatedly, Krieger et al. (2024) found that word predictability from LMs doesn't capture the complete role of contextual information in human sentence processing, particularly with respect to semantic association.

Processing difficulty is commonly indexed using behavioral measures such as reading times, as well as neural measures derived from EEG, including the N400 and the P600 event-related potential (ERP) components. Word predictability has been shown to have robust effects on reading times and the N400 (Kutas and Federmeier, 2011; Ehrlich and Rayner, 1981; Frank et al., 2015; Shain, 2024; Pimentel et al., 2023; Frank and Aumeistere, 2024; Federmeier and Kutas, 1999). In contrast, semantic association between target words and their context has been investigated primarily in ERP studies, with fewer studies examining its relationship to reading times.

ERP studies of semantic association have mostly focused on the N400 component, where semantic association decreases the negative amplitude of the component (Fischler et al., 1983; Kuperberg et al., 2003; Federmeier and Kutas, 1999; Xu et al., 2024; Broderick et al., 2018; Frank and Willems, 2017). However, studies have found that the effect of semantic association on the N400 disappears when there is a delay between the semantically related context and the critical word (Chow et al., 2018; Stone and Rabovsky, 2025). Furthermore, Salicchi and Hsu (2025) found that semantic association didn't explain variance in the N400 component when surprisal was accounted for, while it did in the P600 component, suggesting effects on later processing stages. Evidence of the effect on reading times is less explored. While some studies have found that stronger semantic association decreases reading times (Pynte et al., 2008; Mitchell et al., 2010), other studies indicate that semantic association has no effect on reading times when excluding the variance explained by word predictability (Traxler et al., 2000; Frank, 2017).

Studies of semantic association have mostly relied on stimuli consisting of handcrafted contexts and target words, where they are either semantically similar or not (Federmeier and Kutas, 1999; Fischler et al., 1983; Kuperberg et al., 2003; Stone and Rabovsky, 2025). However, recent studies have attempted to estimate the semantic association using embeddings from LMs (Broderick et al., 2018; Ettinger et al., 2016; Xu et al., 2024; Michaelov et al., 2024; Frank, 2017; Michaelov and Bergen, 2024; Frank and Willems, 2017; Parviz et al., 2011). Thereby, enabling the quantification of semantic association as a continuous measure and facilitating analyses that can extend to naturalistic stimuli.

Embedding-based estimates of semantic association have been conceptualized in a myriad of ways. Firstly, studies deploy different embedding models for extracting the embeddings of the context and the critical word. Most studies use word embeddings, e.g., GloVe, word2vec or fastText (Broderick et al., 2018; Ettinger et al., 2016; Xu et al., 2024; Michaelov et al., 2024; Frank, 2017; Michaelov and Bergen, 2024; Frank and Willems, 2017), however, these models vary in model architecture, embedding size, and training data. Secondly, the context embedding is defined in a variety of ways. Most commonly an average of the word embeddings are used, however, which words are included in the average varies: some studies use all the words in the context (Michaelov and Bergen, 2024; Michaelov et al., 2024; Xu et al., 2024; Broderick et al., 2018), others only content words (Mechtenberg et al., 2025; Frank and Willems, 2017) or a manually select subset of the words (Frank, 2017; Ettinger et al., 2016). Additionally, the length of the context varies. While most studies rely on sentence-level stimuli and use all the preceding words as the context, other studies relying on stimuli consisting of longer text have defined context windows. Frank (2017) defined the context in two separate ways: i) only the sentence preceding the critical word and ii) the four content words immediately preceding the critical word. Similarly, Mechtenberg et al. (2025) examined local and global effects of semantic association by defining context windows of one, two, five, and ten words preceding the critical word, excluding stop words. Finally, different functions for calculating the similarity between the embeddings of the critical word and the context have been employed: While the vast majority utilize the cosine similarity (Ettinger et al., 2016; Xu et al., 2024; Michaelov et al., 2024; Frank, 2017; Michaelov and Bergen, 2024), Pearson's correlation has also been used (e.g., Broderick et al., 2018)

The present study investigated whether semantic association derived from LM embeddings captured aspects of language processing not accounted for by word predictability alone. To accommodate alternative formalizations of semantic association, we defined multiple implementations, varying the embedding model and the size of the contextual window used to compute semantic association. We evaluated these implementations using Bayesian model comparison (Bayes factor) and assess their effects on self-paced reading times and the N400

ERP component. The results of the study showed how the choice of embedding model and the conceptualization of the context can alter the conclusions across neural and behavioral signals.

2. Methods

2.1. Data

The study used data from the Tilburg corpus of Natural Dutch Texts (TiNT; Østergaard et al., 2025). The corpus consists of joined recordings of EEG and self-paced reading (SPR) from 71 participants (whereof 56 participants were included in the analysis of the current study). All participants read eight medium-length (approx. 600 words), natural, Dutch texts of different genres. Seven texts were read using a SPR paradigm, while a single text was read in a rapid serial visual presentation (RSVP) paradigm (the exact text changing from participant to participant). In this study, we only used data recorded during SPR.

Preprocessing of the EEG signal and extraction of ERPs were identical to that of Østergaard et al. (2025). Preprocessing included rereferencing of the electrodes, band-pass filtering, and artifact detection and exclusion. The N400 was defined as the mean amplitude of centroparietal electrodes in the time window 300-500ms after word onset.

2.2. Semantic association

Semantic association was defined as the similarity between the embedding of the context and the embedding of the critical word. Thus, three methodological decisions were required: (1) How to represent the embeddings of the context and the critical word, (2) what context length to use, and (3) which similarity function to apply. In this paper, we defined multiple implementations of semantic association by varying the first two factors, while we used the cosine similarity as the similarity metric across all implementations. Cosine similarity was used, as it is the standard similarity measure for distributional embedding models (Yamada et al., 2020; Reimers and Gurevych, 2019).

(1) Embeddings of the context and the word: Multiple approaches exist for deriving embeddings of text using LMs. Embeddings can be uncontextualized, such as, GloVe, word2vec, or fastText (Pennington et al., 2014; Mikolov et al., 2013; Bojanowski et al., 2017). Such models produce a single embedding for each word in isolation. Alternatively, embeddings can be contextualized. Contextualized embeddings can be derived from transformer models, including both encoders such as BERT (Devlin et al., 2019) and generative models such as GPT and LLama (Radford et al., 2018; Touvron et al., 2023) by retrieving embeddings from the last hidden state. However, embeddings derived directly from pre-trained models typically perform poorly, and thus it has become the norm to adapt contextualized transformer models for embedding tasks, such as semantic text similarity (Reimers and Gurevych, 2019; Gao et al., 2021; Li et al., 2025).

An initial exploration of implementations of semantic association using different embedding models was conducted with simple sentences where the differences in semantic association were handcrafted. The results of the exploration indicated that both contextualized and uncontextualized embedding models were able to differentiate words semantically associated with the context from unrelated words. The results from the models were similar within embedding type (i.e., contextualized or uncontextualized).[2] As such, we selected two candidate models: an uncontextualized word embedding model and a contextualized sentence embedding model.[3]

For the uncontextualized model, we used the word2vec model `wikipedia2vec_nlwiki_20180420_300d`[4] (Yamada et al., 2020). This model was chosen as the training procedure matched previous literature (Broderick et al., 2018; Ettinger et al., 2016; Frank, 2017; Frank and Willems, 2017) and its training data overlaps with the TiNT corpus (Østergaard et al., 2025). As the model only produces one embedding for each word, we used two methods for obtaining the embedding of the context: i) the average of the embeddings of all the words in the context (denoted *WE*), and ii) the average of all the content words in the context (denoted *CWE*). We used the *nl_core_news_sm* model from `spaCy` to extract the part of speech (POS) tags. Content words were identified as words with the POS tags noun, verb, adjective, or adverb.

For the sentence embedding we used `e5-large-trm-nl` (Banar et al., 2025) as it performed well on the Dutch embedding benchmark (MTEB(nld, v1); Banar et al., 2025; Enevoldsen et al., 2025). Sentence embeddings are trained to produce an aggregated embedding over multiple words, thus, it didn't require post-hoc averaging to obtain the embedding of the context. Implementations with sentence embeddings are denoted *SE*.

(2) Context length: Contexts of varying length were defined to examine local and global effects of semantic association. Four distinct context lengths

[2] Results of initial exploration can be found in appendices.

[3] For historical reasons, we call these sentence embeddings, as they initially were trained to embed sentences. However, they have since been expanded to embed entire documents.

[4] Model revisions can be found in appendices.

were used. First, a naive context consisting of all words preceding the critical word was used (*All*). Second, we defined a context consisting of all words in the preceding sentence, as well as in the sentence to which the critical word belonged (*Sentence(N=1)*). Finally, we defined a windowed context, where the context consisted of a fixed number of content words before the target. Here, we used windows of one and two (*Windowed(N=1)* and *Windowed(N=2)*). The windowed implementation was only defined with (content) word embeddings.

In addition to the contexts of different lengths, we defined a weighted average of the word embeddings. The weights followed an exponential forgetting curve, thus, assuming words appearing closer to the critical word were more important. This was implemented as in Equation 1:

$$\text{WE, Weighted} = \sum_{i=1}^{N} 2^{\frac{-i}{4}} \cdot \text{similarity}(w_c, w_i) \quad (1)$$

Here, w_c is the word embedding of the critical word and w_i the word embedding of the word i words away from the critical word. The equation sums over all words preceding the critical word. The denominator (4) determines the half-life of the decay and was chosen such that words at a distance of ten or more from the critical word receive minimal weights. As for the other implementations, the similarity was calculated with the cosine similarity. The weighted average was implemented with word embeddings of all words and word embeddings of content words only. All the implementations of semantic association are summarized in Table 1.

Correlations between semantic association for content words in the corpus extracted using the different implementations are shown in Figure 1. Implementations based on the same type of embedding (i.e., word or sentence embeddings) are strongly correlated, suggesting that they index similar sources of variance. In contrast, correlations across embedding types are substantially weaker.

2.3. Regression models and model comparison

Bayesian hierarchical models were fitted to examine the effect of the different implementations of semantic association on the two dependent variables: self-paced reading times and the N400. The models were fitted in Stan (version 2.32.2; Stan Development Team, 2023) using the `brms` package (version 2.22.0; Bürkner, 2017) in R (R Core Team, 2024). All predictors were z-score standardized. Words with reading times lower than 100 ms or greater than 3000 ms were excluded from analysis. Only content words (i.e., nouns, verbs, adjectives,

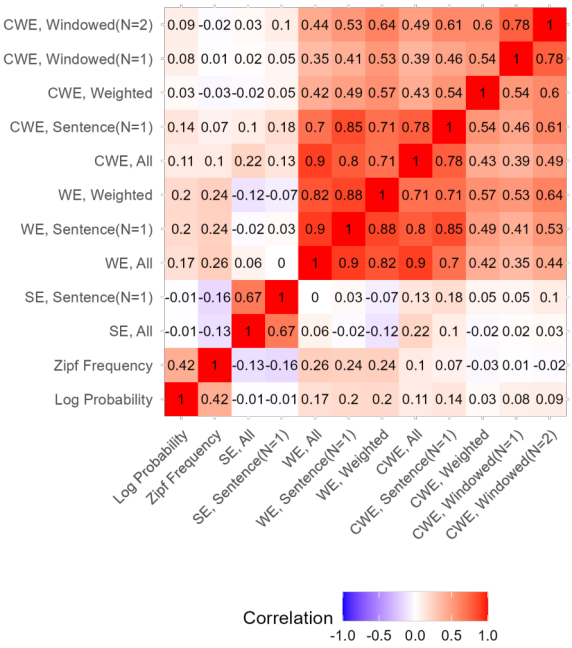

Figure 1: Pearson's correlation coefficients between implementations of semantic association, log-probability of words, and Zipf word frequency for all content words in the corpus.

and adverbs) were included in the analysis. As embeddings extracted from the word2vec model only exist for a finite number of words, the data loss slightly differed across different implementations of semantic association.[5] The models were fitted on complete cases across all implementations.

The models were run with two predictors, log-probability lp, estimated by the average word probability from four GPT models (see Østergaard et al., 2025), and semantic association sem. One regression model for each of the two dependent variables and each of the 10 implementations of semantic association was fitted, resulting in 20 separate models. Uncorrelated group-level intercepts and slopes for both lp and sem were estimated for each participant, document, and word. For the reading time (RT) model, a log-normal likelihood was used, while for the N400 model, a Gaussian likelihood was used (See Equations 2, 3, and 4).

$$\text{RT} \sim LogNormal(\mu, \sigma) \quad (2)$$

$$\text{N400} \sim Normal(\mu, \sigma) \quad (3)$$

$$\begin{aligned} \mu =& \alpha + u_{participant,0} + u_{document,0} + \\ & u_{word,0} + (\beta_1 + u_{participant,1} + \\ & u_{document,1} + u_{word,1}) \cdot lp + (\beta_2 + \\ & u_{participant,2} + u_{document,2} + u_{word,2}) \cdot sem \end{aligned} \quad (4)$$

[5] Data loss across all implementations is reported in appendices.

Name	Embedding model	Context	Words
SE, All	Sentence Embeddings	All preceding words	All
SE, Sentence(N=1)	Sentence Embeddings	One sentence before the target sentence	All
WE, All	Word Embeddings	All preceding words	All
WE, Sentence(N=1)	Word Embeddings	One sentence before the target sentence	All
WE, Weighted	Word Embeddings	All preceding words (weighted)	All
CWE, All	Word Embeddings	All preceding words	Content words
CWE, Sentence(N=1)	Word Embeddings	One sentence before the target sentence	Content words
CWE, Weighted	Word Embeddings	All preceding words (weighted)	Content words
CWE, Windowed(N=2)	Word Embeddings	One content word preceding the target	Content words
CWE, Windowed(N=2)	Word Embeddings	Two content word preceding the target	Content words

Table 1: All implementations of semantic association used in the current study. The implementations will be referred to by the name in the name column.

Different priors were used for the reading times and the N400 model, as the scales of the dependent variables were different, i.e., reading times in ms and ERP components in μV. For all models, regularizing priors were used to ensure stable and plausible estimates (Nicenboim et al., 2025). The priors for the reading times model were as follows:

$$\begin{aligned}\alpha &\sim Normal(5.5, 1)\\ \beta &\sim Normal(0, .1)\\ u &\sim Normal(0, sd)\\ sd &\sim Normal_+(0, .5)\\ \sigma &\sim Normal_+(0, .5)\end{aligned}$$

The priors for the models of the ERP components were:

$$\begin{aligned}\alpha &\sim Normal(0, 20)\\ \beta &\sim Normal(0, 10)\\ u &\sim Normal(0, sd)\\ sd &\sim Normal_+(0, 10)\\ \sigma &\sim Normal_+(0, 10)\end{aligned}$$

To assess the influence of the various implementations of semantic association on reading comprehension (i.e., reading times and the N400), we used Bayes factors. Bayes factor provides a framework for Bayesian hypothesis testing by quantifying evidence in favor of a model (M_0) given another (M_1). This is calculated as the ratio between the marginal likelihoods of the two models, which in turn responds to two hypotheses. (see Equation 5).

$$BF_{01} = \frac{p(y|M_0)}{p(y|M_1)} \tag{5}$$

As such, a Bayes factor of one indicates no evidence for either model, a Bayes factor of 10 would be strong evidence for M_0, and a Bayes factor of $1/10$ indicates strong evidence for M_1. We used the Savage-Dickey density ratio method to calculate Bayes factor, as it provides a convenient method for computing Bayes factor for nested models with a point null hypothesis (Dickey and Lientz, 1970). We specifically tested the null hypothesis that there's no effect of semantic association in the models when log-probability is included. The Savage-Dickey ratio was calculated separately for each model as in Equation 6.

$$BF_{01} = \frac{p(\beta_2 = 0|y)}{p(\beta_2 = 0)} \tag{6}$$

Here, y denotes the observed data and β_2 is the coefficient for semantic association.

As Bayes factor is sensitive to the choice of prior, we conducted a sensitivity analysis by varying the width of the prior for β_2 while keeping the priors for all other parameters fixed. For the reading times models, we used additional priors of $\beta_2 \sim Normal(0, .05)$ and $\beta_2 \sim Normal(0, .5)$, and for the N400 models, $\beta_2 \sim Normal(0, 1)$ and $\beta_2 \sim Normal(0, 2)$. For a more elaborate explanation of Bayes factor and the Savage-Dickey ratio, see Nicenboim and Vasishth (2016).

Most models were fitted using four chains with 2,000 iterations, where half the iterations were warm-up samples. However, six models required 3,000 iterations to ensure stable posterior sampling. The models reported in this paper had no divergent transitions, $\hat{R}s \leq 1.03$, and the number of bulk and tail effective samples was at least 119 and an average of 1,477.

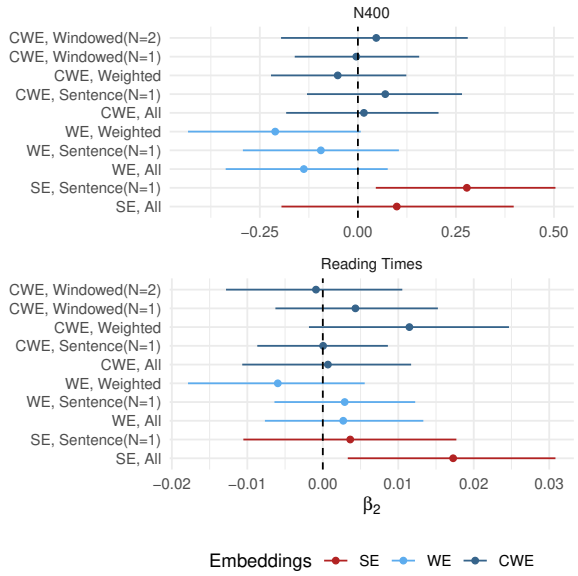

Figure 2: Regression coefficients and 95% credible intervals for semantic association β_2 as estimated by the different implementations. Every point represents β_2 from a separate regression model. The models with N400 as the dependent variable are measured in μV, while the reading times models are measured in ms (thus, the scales of the x-axis differ across the two).

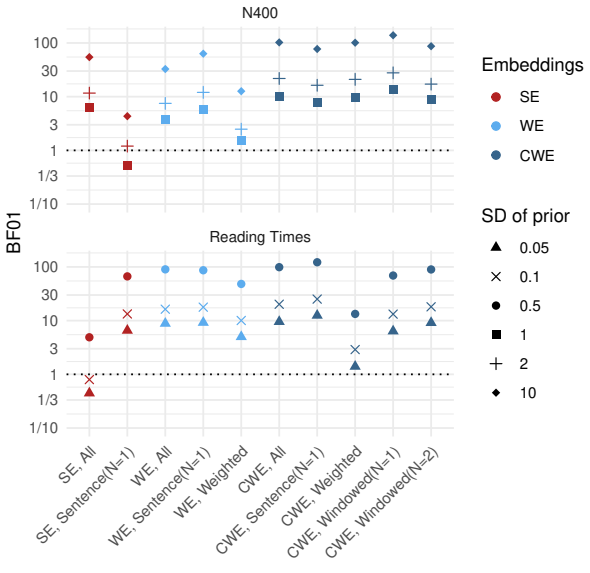

Figure 3: Bayes factor. $BF01 > 1$ indicates more evidence for the null hypothesis (i.e., no effect of semantic association) and $BF01 < 1$ indicates more evidence for the alternative hypothesis (i.e., an effect of semantic association). Each Bayes factor was calculated for separately fitted models with different standard deviations (SD) for the prior of β_2.

3. Results

Figure 2 displays the coefficients for semantic association (β_2 in Equation 4) estimated by the 20 regression models using the values from the different implementations of semantic association to predict the N400 and self-paced reading times. Bayes factor for β_2 across the regression models reported in Figure 3. The results from the Bayes factor showed anecdotal evidence ($BF01 \in \{1, 1/3\}$) for an effect of semantic association in only two models (*SE, Sentence(N=1)* for the N400 and *SE, All* for reading times). Across the rest of the models for both dependent variables, there was the most evidence for the null hypothesis, i.e., no effect of semantic association.

Embedding models: The results of the regression models indicate that the choice of embedding model when calculating semantic association impacts the estimated effects on neural and behavioral measures. This pattern was particularly pronounced for the models of the N400. The estimated effect of semantic association on the N400 when using sentence embeddings (*SE*) was positive, meaning that less semantically associated words elicited a more negative N400 amplitude, consistent with previous literature (Fischler et al., 1983; Kuperberg et al., 2003; Federmeier and Kutas, 1999; Xu et al., 2024; Broderick et al., 2018; Frank and Willems, 2017). In contrast, when semantic association was calculated using word embeddings (*WE*), the direction of the effect reversed, i.e., a negative estimate. When semantic association was calculated using the same word embeddings but only embeddings of the content words in the context (*CWE*), the estimated effect of semantic association was close to zero. For reading times, only the model of semantic association from the *SE, All* implementation indicated an effect. This model estimated a positive effect of semantic association on reading times; thus, reading times increased when words were more semantically associated to the context. The estimated coefficients for semantic association for the rest of the models were smaller and generally close to zero.

Context length: The results show that the length of the context matters only for the semantic association defined with sentence embeddings. The implementations of semantic association using word embeddings (both *WE* and *CWE*) showed similar effects on both the N400 and reading times across all contexts (*All*, *Sentence*, *Weighted*, and *Windowed*). For the implementations relying on sentence embeddings, the effect of context appeared to play a more substantial role. On the N400, the effect of semantic association was largest for the regression model with *SE, Sentence(N=1)*, while the largest effect of semantic association on reading

times was estimated by the model with *SE, All*.

4. Discussion

In this study, we employed both uncontextualized word embeddings and contextualized sentence embeddings to estimate semantic association. These two embedding types appear to capture distinct patterns in the text. This is both apparent from the correlations between the different implementations of semantic association (Figure 1) but, more importantly, for the estimated effects of semantic association on self-paced reading times and the N400 too (Figure 2 and Figure 3). The results of the regression models show that the type of embeddings used influences the estimated effect of semantic association. This finding is most pronounced for the N400, where the estimated effect reverses direction depending on whether semantic association is computed using sentence or word embeddings. A positive effect of semantic association on the N400 is found when using sentence embeddings, while the opposite (i.e., a negative effect) is estimated with word embedding-based semantic association. In contrast, previous literature using similar uncontextualized word embeddings to calculate semantic association reports a positive effect of semantic association on the N400 (Broderick et al., 2018; Frank and Willems, 2017; Xu et al., 2024). It is important to note that the Bayes factor indicated no evidence for the negative effects estimated for word embedding-based semantic association, however, anecdotal evidence for one of the models with a positive effect for semantic association from sentence embeddings.

What could be possible explanations for the observed difference in semantic association when computed with different types of embedding models? One important distinction between sentence embeddings and word embeddings when using them to create context representations lies in how information was aggregated. Although sentence embeddings output an aggregation of embeddings too, the model has been trained to produce a semantically coherent representation in which more informative words receive greater weight. In this study, the implementations of semantic association using word embeddings from word2vec relied on a naive approach, where an unweighted average was utilized, inspired by previous approaches (Michaelov and Bergen, 2024; Michaelov et al., 2024; Xu et al., 2024; Broderick et al., 2018; Mechtenberg et al., 2025; Frank and Willems, 2017; Frank, 2017; Ettinger et al., 2016). As such, important information could be lost in the context representations derived from the word embedding implementations — especially for the implementations of longer contexts (i.e., *WE, All* and *WE, Sentence(N=1)*). Studies finding positive effects of word embedding-based semantic association on the N400 have generally relied on shorter contexts (either because sentence-level stimuli were used or because they defined short context lengths), thus minimizing the information loss when averaging over embeddings. This speculation is supported by our initial exploration, where both sentence and word embeddings produced effects in the same direction using sentence pairs from Federmeier and Kutas (1999).[6] While the weighted implementations of semantic association (*WE, Weighted* and *CWE, Weighted*) were cognitively motivated implementations, discounting the influence of word embeddings on the overall average based an exponential forgetting curve, this approach didn't seem promising given the results of the current study. The weights were solely based on distances to the critical word, thus words were not weighted based on their semantic relevance.

To our knowledge, no previous works have used sentence embeddings for studying semantic association in sentence processing. The present findings suggest that this approach provides a promising method for estimating semantic relations between contexts and target words. Implementations based on sentence embeddings showed the most reliable effects on both the N400 and self-paced reading times, as reflected by the size of the regression coefficients and Bayes factors. Post-hoc qualitative analyses further indicated that sentence embeddings are more sensitive to the general themes of the texts compared to averaged word embeddings. Figure 4 illustrates a difference between sentence embeddings and word embeddings for calculating semantic association in two examples of sentence pairs from the corpus. In the first pair, the word "dragon" appears in both sentences. Only the *SE, All* implementation captures the association between "draak" (English: "dragon") and the story "Mijn Heer Zak met Rijst"[7] in the first sentence, while *SE, Sentence (N=1)* detects the association when the word reappears in the second sentence. In contrast, none of the word embedding implementations indicate a strong association for "dragon" in either instance. In sentence pair B from the text "Nomadisch pastoralisme", a similar pattern is observed with the two words "Nomadisme" (English: "Nomadism") and "nomadische" (English: "nomadic"). While these examples were selected for illustrative purposes, they suggest that sentence embeddings capture a thematically coherent representation of semantic association in natural texts.

Naturally, this interpretation depends on the operationalization of semantic association. In the

[6] Results of initial exploration can be found in appendices

[7] A fairy tale about a Dragon King.

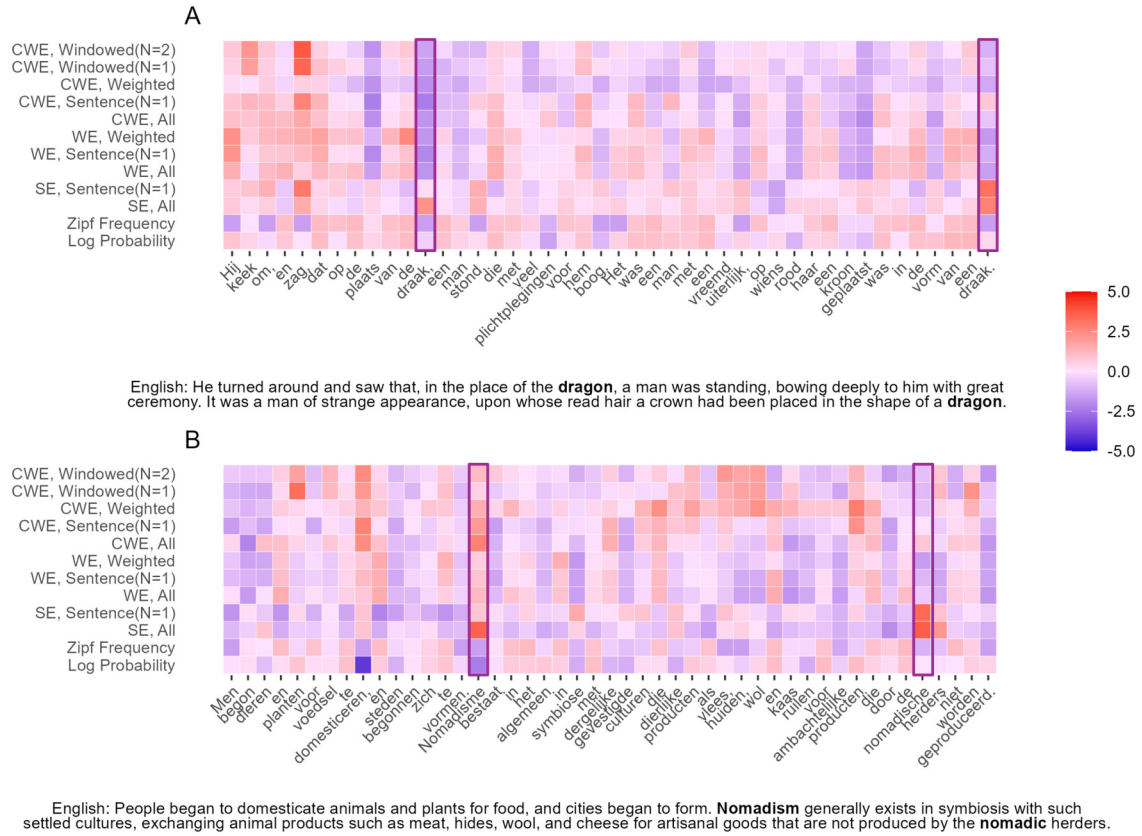

Figure 4: Semantic association (as estimated by different implementations), log-probability, and Zipf frequency of words in two sentence pairs from two different documents in the corpus. Highlighted are the words “draak” (English: “dragon”) in sentences A and “Nomadisme” (English: “Nomadism”) and “nomadische” (English: “nomadic”) in sentences B. All variables (i.e., log probability, Zipf frequency, and semantic associations) are z-score standardized.

present study, semantic association was defined as the similarity between embeddings of the context and the critical word, following prior work (Broderick et al., 2018; Ettinger et al., 2016; Xu et al., 2024; Michaelov et al., 2024; Frank, 2017; Michaelov and Bergen, 2024; Frank and Willems, 2017). This operationalization assumes that the embeddings encode multiple aspects of meaning, including both shared features (e.g., *nurse* and *mechanic* as occupations) and thematic relations (e.g., *nurse* and *hospital* as related). Consequently, embedding similarity captures similarities in the features of the word as well as their relatedness. Following this definition, repeated words will inflate semantic association, as the similarity between two identical embeddings is one (i.e., maximum semantic association). However, this property applies for all the implementations of semantic association considered in the current work, thus, the effect of repetition can’t account for the differences observed in example A in Figure 4.

The results of the current study are exploratory, and further work is required to identify under which conditions specific implementations of LM embedding-based estimates of semantic association differ. The analysis was based exclusively on texts from a single corpus (TiNT; Østergaard et al., 2025), which consists of medium-length, Dutch texts. Not only does this corpus stand in contrast to previously used stimuli by the length of the texts (as touched upon above), but also in the language. As Dutch has been less extensively studied than English, the quality of the embedding models may differ, potentially affecting their performance. As such, semantic association as estimated by the different implementations in this study should be validated on other corpora to determine whether it is possible to replicate previously reported effects of semantic association.

The most prominent finding of this study lies in the importance of the embedding model for estimating semantic association. Only one word embedding model and one sentence embedding model were included in the analysis, as initial explorations indicated minimal differences between models within each embedding type. However, in light of the results of the current study, further ex-

ploration of different embedding models would be interesting.

5. Conclusion

This study examined the effects of LM embedding-based semantic association on the self-paced reading of medium-length, Dutch texts. The findings demonstrate that the conclusions critically depend on how semantic association is implemented, particularly with respect to the embedding model. While uncontextualized word embeddings (e.g., word2vec) have previously been used to examine semantic association in sentence processing and showed effects on the N400 (Xu et al., 2024; Broderick et al., 2018; Frank and Willems, 2017), we observed no effects on either the N400 or self-paced reading times. In contrast, semantic association estimated with sentence embeddings was found to be predictive of processing difficulty. These results suggest sentence embeddings to be a promising approach for examining the effects of semantic association in natural reading.

6. Code availability

Code for reproducing the analysis is publicly available from GitHub at `https://github.com/saraoe/semantic_association`.

7. Ethics Statement

The study utilized human data from the Tilburg corpus of Natural Dutch Texts (TiNT) collected by Østergaard et al. (2025). The data has received an ethics approval and is licensed under a CC-BY-NC-SA license.

8. Bibliographical References

Christoph Aurnhammer, Francesca Delogu, Harm Brouwer, and Matthew W. Crocker. 2023. The P600 as a continuous index of integration effort. *Psychophysiology*, 60(9):e14302.

Nikolay Banar, Ehsan Lotfi, Jens Van Nooten, Cristina Arhiliuc, Marija Kliocaite, and Walter Daelemans. 2025. Mteb-nl and e5-nl: Embedding benchmark and models for dutch.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Michael P. Broderick, Andrew J. Anderson, Giovanni M. Di Liberto, Michael J. Crosse, and Edmund C. Lalor. 2018. Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. *Current Biology*, 28(5):803–809.e3.

Harm Brouwer, Hartmut Fitz, and John Hoeks. 2012. Getting real about Semantic Illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research*, 1446:127–143.

Nyssa Z. Bulkes, Kiel Christianson, and Darren Tanner. 2020. Semantic constraint, reading control, and the granularity of form-based expectations during semantic processing: Evidence from ERPs. *Neuropsychologia*, 137:107294.

Paul-Christian Bürkner. 2017. brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1):1–28.

Wing-Yee Chow, Ellen Lau, Suiping Wang, and Colin Phillips. 2018. Wait a second! delayed impact of argument roles on on-line verb prediction. *Language, Cognition and Neuroscience*, 33(7):803–828.

Michael Dambacher, Reinhold Kliegl, Markus Hofmann, and Arthur M. Jacobs. 2006. Frequency and predictability effects on event-related potentials during reading. *Brain Research*, 1084(1):89–103.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

James M. Dickey and B. P. Lientz. 1970. The Weighted Likelihood Ratio, Sharp Hypotheses about Chances, the Order of a Markov Chain. *The Annals of Mathematical Statistics*, 41(1):214–226.

Susan F. Ehrlich and Keith Rayner. 1981. Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, 20(6):641–655.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua,

Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Mariya Hendriksen, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Šuppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri Krishnakumar, Anna Maksimova, Silvan Wehrli, Maria Tikhonova, Henil Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Hongjin Su, Jimmy Lin, Howard Yen, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. 2025. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*.

Allyson Ettinger, Naomi Feldman, Philip Resnik, and Colin Phillips. 2016. Modeling N400 amplitude using vector space models of word representation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 38(0).

Kara D. Federmeier and Marta Kutas. 1999. A Rose by Any Other Name: Long-Term Memory Structure and Sentence Processing. *Journal of Memory and Language*, 41(4):469–495.

I. Fischler, P. A. Bloom, D. G. Childers, S. E. Roucos, and N. W. Perry. 1983. Brain potentials related to stages of sentence verification. *Psychophysiology*, 20(4):400–409.

Stefan L. Frank. 2017. Word Embedding Distance Does not Predict Word Reading Time. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 39(0).

Stefan L. Frank and Anna Aumeistere. 2024. An eye-tracking-with-EEG coregistration corpus of narrative sentences. *Language Resources and Evaluation*, 58(2):641–657.

Stefan L. Frank, Leun J. Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11.

Stefan L. Frank and Roel M. Willems. 2017. Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, 32(9):1192–1203.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Hale. 2001. A Probabilistic Earley Parser as a Psycholinguistic Model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Benedict Krieger, Harm Brouwer, Christoph Aurnhammer, and Matthew W. Crocker. 2024. On the limits of LLM surprisal as functional Explanation of ERPs. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46(0).

Gina R. Kuperberg, Tatiana Sitnikova, David Caplan, and Phillip J. Holcomb. 2003. Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Brain Research. Cognitive Brain Research*, 17(1):117–129.

Marta Kutas and Kara D. Federmeier. 2011. Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, 62(1):621–647.

Litao Li, Leo Song, Steven Ding, Benjamin C. M. Fung, and Philippe Charland. 2025. Transforming Generic Coder LLMs to Effective Binary Code Embedding Models for Similarity Detection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Steven G. Luke and Kiel Christianson. 2018. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2):826–833.

Hannah Mechtenberg, James Reilly, Emily B Myers, and Jonathan E Peelle. 2025. Measuring brain sensitivity to semantic distance in spoken narrative comprehension.

James A. Michaelov, Megan D. Bardolph, Cyma K. Van Petten, Benjamin K. Bergen, and Seana

Coulson. 2024. Strong Prediction: Language Model Surprisal Explains Multiple N400 Effects. *Neurobiology of Language*, 5(1):107–135.

James A. Michaelov and Benjamin K. Bergen. 2024. On the Mathematical Relationship Between Contextual Probability and N400 Amplitude. *Open Mind*, 8:859–897.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space.

Jeff Mitchell, Mirella Lapata, Vera Demberg, and Frank Keller. 2010. Syntactic and Semantic Factors in Processing Difficulty: An Integrated Measure. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 196–206, Uppsala, Sweden. Association for Computational Linguistics.

Bruno Nicenboim, Daniel J. Schad, and Shravan Vasishth. 2025. The influence of priors: sensitivity analysis. In *Introduction to Bayesian Data Analysis for Cognitive Science*, 1st edn edition, chapter 3.4, page 634. Chapman & Hall.

Bruno Nicenboim and Shravan Vasishth. 2016. Statistical methods for linguistic research: Foundational Ideas—Part II. *Language and Linguistics Compass*, 10(11):591–613.

Mante S. Nieuwland and Jos J. A. Van Berkum. 2005. Testing the limits of the semantic illusion phenomenon: ERPs reveal temporary semantic change deafness in discourse comprehension. *Cognitive Brain Research*, 24(3):691–701.

OpenAI. 2025. GPT-5.2 via OpenAI API.

Mehdi Parviz, Mark Johnson, Blake Johnson, and Jon Brock. 2011. Using Language Models and Latent Semantic Analysis to Characterise the N400m Neural Response. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 38–46, Canberra, Australia.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Tiago Pimentel, Clara Meister, Ethan G. Wilcox, Roger Levy, and Ryan Cotterell. 2023. On the Effect of Anticipation on Reading Times.

Joel Pynte, Boris New, and Alan Kennedy. 2008. On-line contextual influences during reading normal text: A multiple-regression analysis. *Vision Research*, 48(21):2172–2183.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. *OpenAI Technical Report*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.

Lavinia Salicchi and Yu-Yin Hsu. 2025. Different Reading Processing Stages or Different Brain Areas? A Computational Cognitive Investigation on N400, P600, and PNP. In *Computational Psycholinguistics Meeting 2025*. Conference presentation abstract.

Cory Shain. 2024. Word Frequency and Predictability Dissociate in Naturalistic Reading. *Open Mind*, 8:177–201.

Stan Development Team. 2023. Stan Reference Manual.

Kate Stone and Milena Rabovsky. 2025. The Role of Syntactic and Semantic Cues in Preventing Temporary Illusions of Plausibility. *Journal of Cognitive Neuroscience*, 37(9):1535–1561.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models.

Matthew J. Traxler, Donald J. Foss, Rachel E. Seely, Barbara Kaup, and Robin K. Morris. 2000. Priming in Sentence Processing: Intralexical Spreading Activation, Schemas, and Situation Models. *Journal of Psycholinguistic Research*, 29(6):581–595.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Roslyn Wong, Erik D. Reichle, and Aaron Veldre. 2024. Prediction in reading: A review of predictability effects, their theoretical implications, and beyond. *Psychonomic Bulletin & Review*.

Haoyin Xu, Masaki Nakanishi, and Seana Coulson. 2024. Revisiting Joke Comprehension with Surprisal and Contextual Similarity: Implication from N400 and P600 Components. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46(0).

Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. 2020. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 23–30. Association for Computational Linguistics.

Sara Møller Østergaard, Lenneke Lichtenberg, Laura Boon, and Bruno Nicenboim. 2025. A Corpus of Joint EEG and Self-Paced Reading of Natural Dutch Texts.

A. Appendices

A.1. Hugging Face References and Revision

Hugging Face Reference	Revision
Word2vec/wikipedia2vec_nlwiki_20180420_300d (Yamada et al., 2020)	f7c83ecdf955a4f482a12517ca52a1f4b81e43cf
clips/e5-large-trm-nl (Banar et al., 2025)	683333f86ed9eb3699b5567f0fdabeb958d412b0
Word2vec/wikipedia2vec_enwiki_20180420_100d (Yamada et al., 2020)	7e4d6d224b95a5c351e2a47232701c4403ffbc16
fse/word2vec-google-news-300	528f381952a0b7d777bb4a611c4a43f588d48994
sentence-transformers/all-MiniLM-L6-v2	c9745ed1d9f207416be6d2e6f8de32d1f16199bf
intfloat/multilingual-e5-large (Wang et al., 2024)	0dc5580a448e4284468b8909bae50fa925907bc5
spacy/nl_core_news_sm (Honnibal et al., 2020)	b9d28fe480eeacf9809fbd5ead5ef1ff27d9394e

Table 2: Overview of Hugging Face models used for the analysis.

A.2. Data loss

The data loss for the different implementations of semantic association, i.e., the number of words in the corpus for which semantic association could not be calculated. The use of sentence embeddings resulted in a lower data loss compared to the word embedding implementations, as word embeddings extracted from a word2vec model only exist for a finite number of words. The data loss is largest for the *CWE, Windowed(N=1)*.

Implementation	Data loss	
	Overall	Content words
SE, All	0.17%	0.08%
SE, Sentence(N=1)	0.17%	0.08%
WE, All	1.78%	2.06%
WE, Sentence(N=1)	1.78%	2.06%
WE, Weighted	1.78%	2.06%
CWE, All	2.09%	2.33%
CWE, Sentence(N=1)	2.17%	2.38%
CWE, Weighted	2.09%	2.33%
CWE, Windowed(N=1)	3.91%	4.25%
CWE, Windowed(N=2)	2.21%	2.39%

Table 3: Data loss caused by the different implementations of semantic association. The table shows the overall data loss across all words in the corpus, and the data loss for the current analysis that only considers content words.

A.3. Validation of semantic association

To validate the implementations of semantic association, we used data from Federmeier and Kutas (1999). As seen in Ettinger et al. (2016), we wanted to validate that the model identified expected targets as more semantically similar to the context compared to the within-category and between-category target words. In addition to the original stimuli, we added an unrelated target word (identical to an expected target in another context). Moreover, using OpenAI's GPT-5.2 (OpenAI, 2025), we generated longer contexts of approx. 100 words to validate the models on contexts longer than the two-sentence context provided in the original data. The results of the validation are shown in Figure 5. All the implementations identify the correct ordering of semantic association between the target words and the contexts. The sentence embedding models, `all-MiniLM-L6-v2` and `intfloat/multilingual-e5-large`, generally exhibit less variance overlap between target words (and most notably between the expected and unexpected target) as compared to the word embedding models, `enwiki_20180420_100d` and `word2vec-google-news-300`.

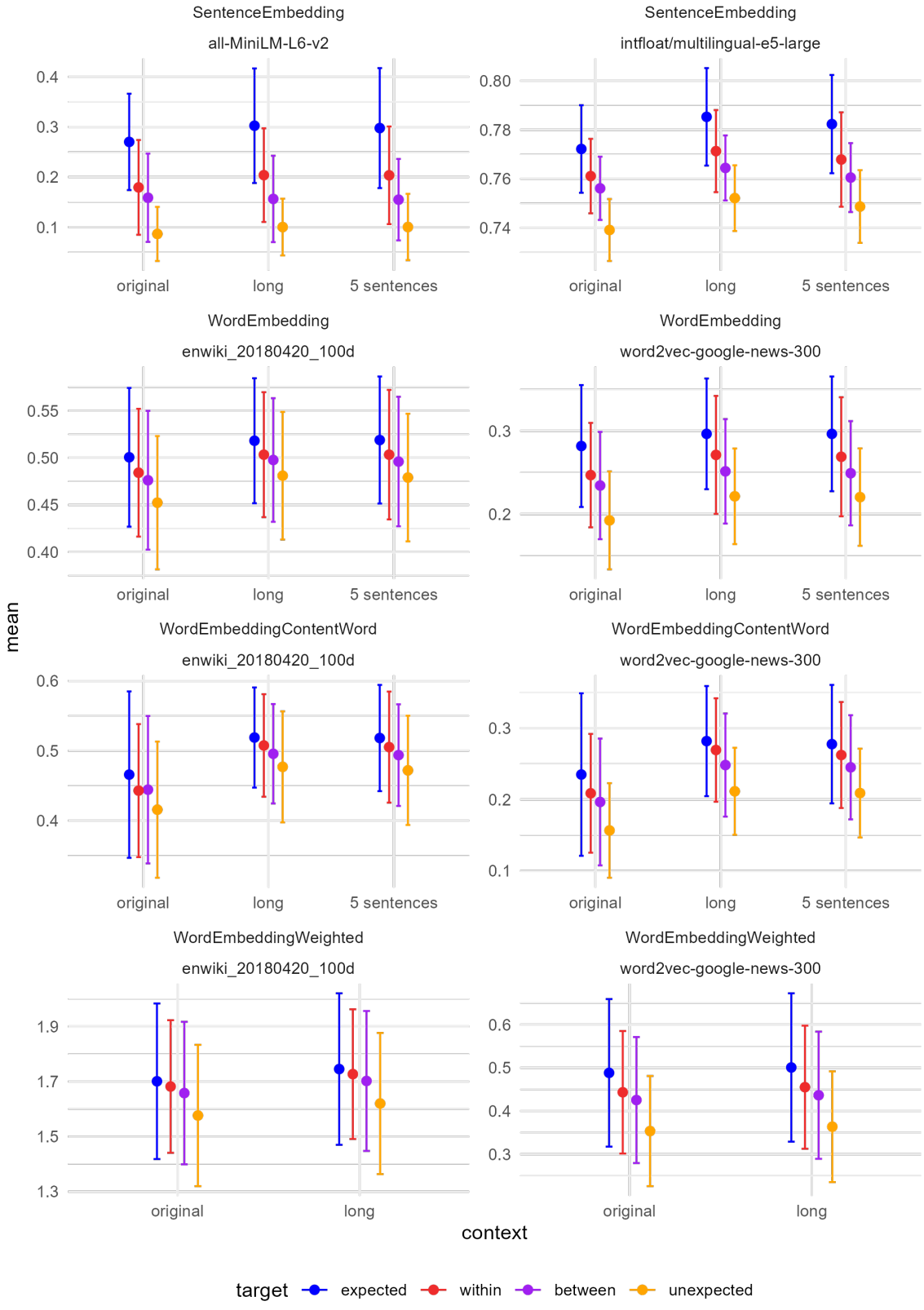

Figure 5: Average semantic association between context and target words (expected, within, between, and unexpected) from Federmeier and Kutas (1999) given different implementations of semantic association. The plot is divided by the embedding model and the implementation of semantic association. The x-axis indicates the context the target was associated with, where "long" means the original and the generated longer context, "original" means the original context, "5 sentences" means the two sentences in the original context and three more from the longer context. The error bars indicate the standard deviation. Note that the y-axes are different across plots.