# Can Peanuts Fall in Love with Distributional Semantics?

**James A. Michaelov (j1michae@ucsd.edu)**
**Seana Coulson (scoulson@ucsd.edu)**
**Benjamin K. Bergen (bkbergen@ucsd.edu)**
Department of Cognitive Science, University of California, San Diego
9500 Gilman Dr, La Jolla, CA 92093, USA

## Abstract

Context changes expectations about upcoming words—following a story involving an anthropomorphic peanut, comprehenders expect the sentence *the peanut was in love* more than *the peanut was salted*, as indexed by N400 amplitude (Nieuwland & van Berkum, 2006). This updating of expectations has been explained using Situation Models—mental representations of a described event. However, recent work showing that N400 amplitude is predictable from distributional information alone raises the question whether situation models are necessary for these contextual effects. We model the results of Nieuwland and van Berkum (2006) using six computational language models and three sets of word vectors, none of which have explicit situation models or semantic grounding. We find that a subset of these can fully model the effect found by Nieuwland and van Berkum (2006). Thus, at least some processing effects normally explained through situation models may not in fact require explicit situation models.

**Keywords:** psycholinguistics; human language comprehension; event-related brain potentials; N400; natural language processing; deep learning; language models; word vectors

## Introduction

It is widely believed that prediction plays a key role in language processing, with more predictable words being processed more easily (Fischler & Bloom, 1979; Kutas & Hillyard, 1984; Levy, 2008; Kutas, DeLong, & Smith, 2011; Van Petten & Luka, 2012; DeLong, Troyer, & Kutas, 2014; Luke & Christianson, 2016; Kuperberg, Brothers, & Wlotko, 2020). Perhaps the strongest evidence for this comes from the N400, a neural signal of processing difficulty that is highly correlated with lexical probability—contextually probable words elicit an N400 response of smaller (less negative) amplitude than contextually improbable words, whether predictability is determined based on human judgements (Kutas & Hillyard, 1984; for review see Van Petten & Luka, 2012) or a corpus (Parviz, Johnson, Johnson, & Brock, 2011; Frank, Otten, Galli, & Vigliocco, 2015; Aurnhammer & Frank, 2019b; Merkx & Frank, 2021; Szewczyk & Federmeier, 2022; Michaelov, Coulson, & Bergen, 2022; Michaelov, Bardolph, Van Petten, Bergen, & Coulson, 2023).

A striking feature of the predictions indexed by the N400 is how flexible they can be. Under normal circumstances, a sentence such as *the peanut was in love* would be highly improbable, much more so than *the peanut was salted*. Following the short story in (1), however, this changes (Nieuwland & van Berkum, 2006).

(1) A woman saw a dancing peanut who had a big smile on his face. The peanut was singing about a girl he had just met. And judging from the song, the peanut was totally crazy about her. The woman thought it was really cute to see the peanut singing and dancing like that.

In fact, Nieuwland and van Berkum (2006), who tested this in Dutch, found that in the context of (1), the last word of *de pinda was **verliefd*** ('the peanut was **in love**') elicited a smaller N400 than *de pinda was **gezouten*** ('the peanut was **salted**'). How does such a dramatic reversal occur?

One possibility put forward by Nieuwland and van Berkum (2006) is that while reading the context, the reader's mental representation of the peanut is altered such that it is treated as an animate entity. This, as Nieuwland and van Berkum (2006) note, is in line with theories of situation models, which argue that we track the entities under discussion, as well as their properties and relations. Such accounts generally involve explicit structures or schemata, grounding in world knowledge or experience, extraction of propositional information, or a combination of these (see, e.g., Bransford, Barclay, & Franks, 1972; Kintsch & van Dijk, 1978; Johnson-Laird, 1980; Garnham, 1981; Johnson-Laird, 1983; van Dijk & Kintsch, 1983; Kintsch, 1988; Zwaan, Langston, & Graesser, 1995; Zwaan, Magliano, & Graesser, 1995; Radvansky, Zwaan, Federico, & Franklin, 1998; Kintsch, 1998; Zwaan & Radvansky, 1998; Zwaan & Madden, 2004; Kintsch, 2005; Van Berkum, Koornneef, Otten, & Nieuwland, 2007; Kintsch & Mangalath, 2011; Butcher & Kintsch, 2012; Zwaan, 2014, 2016; Zacks & Ferstl, 2016; Kintsch, 2018; Hoeben Mannaert & Dijkstra, 2021). On a situation model account, the reader alters their semantic representation of the peanut to give it animate features in accordance with the information that it can sing, dance, and show emotions, thereby facilitating the processing of *in love*.

The hypothesis that structured or grounded situation models explain N400 effects such as those found by Nieuwland and van Berkum (2006) is generally accepted (e.g., Hagoort & van Berkum, 2007; Filik & Leuthold, 2008; Warren, McConnell, & Rayner, 2008; Rosenbach, 2008; Ferguson & Sanford, 2008; Ferguson, Sanford, & Leuthold, 2008; Menenti, Petersson, Scheeringa, & Hagoort, 2009; Bicknell, Elman, Hare, McRae, & Kutas, 2010; de Groot, 2011; Metusalem et al., 2012; Aravena et al., 2014; Zwaan, 2014; Xi-

ang & Kuperberg, 2015; Kuperberg et al., 2020) and has been shown to be viable using computational models (Venhuizen, Crocker, & Brouwer, 2019). However, there are alternative explanations.

The present study asks whether the effect can instead be explained by lexical preactivation based on distributional linguistic knowledge, following the findings that the statistics of language can be used to model N400 effects (Ettinger, Feldman, Resnik, & Phillips, 2016; Michaelov & Bergen, 2020; Michaelov, Bardolph, Coulson, & Bergen, 2021; Michaelov & Bergen, 2022a; Uchida, Lair, Ishiguro, & Dominey, 2021; Michaelov et al., 2023) and predict single-trial N400 amplitude (Chwilla & Kolk, 2005; Parviz et al., 2011; Van Petten, 2014; Frank et al., 2015; Aurnhammer & Frank, 2019a, 2019b; Merkx & Frank, 2021; Michaelov et al., 2021; Szewczyk & Federmeier, 2022; Michaelov et al., 2023).

Specifically, we look at two possible ways in which this might arise. One, which we refer to as *event-level priming*, refers to the idea that a word associated with a previously-discussed event may be more likely to be predicted by virtue of this. This is something that has been previously reported in the N400—Metusalem et al. (2012), for example, found that that merely being related to the event under discussion leads to a smaller N400 response to a word even when that word is inappropriate. Michaelov and Bergen (2022a) model this with transformer language models—systems trained to calculate the probability of a word given its context based on the statistics of language alone—showing that this effect is explainable with distributional information. Thus, it may be the case that the fact that *in love* is related to, for example, being *crazy about* someone that leads to it being predicted to be more likely than *salted*. Following Michaelov and Bergen (2022a), we investigate this using 6 Dutch transformer language models (Havinga, 2021, 2022a, 2022b, 2022c; de Vries et al., 2019; Delobelle, Winters, & Berendt, 2020), testing whether they show the same effect as humans—that is, whether they predict the canonical sentence the *peanut was salted* to be less likely than the noncanonical sentence *the peanut was in love*.

An alternative possibility is *lexical priming*. More simply than in the case of event-level priming, it may be the case that the preceding context involving words such as *dancing*, *smile*, *singing*, *crazy*, and *cute* exerts a stronger pressure on prediction of *in love* than *peanut* does on *salted*. Intuitively, one might expect that a system (neurocognitive or computational) displaying event-level priming is likely to display lexical priming—indeed, lexical priming is a possible mechanism by which at least some part of event-level priming could be achieved. The fact that lexical priming is likely to occur in a system displaying event-level priming is also supported by the fact that language models show both (Kassner & Schütze, 2020; Misra, Ettinger, & Rayz, 2020; Michaelov & Bergen, 2022a). Thus, in the present study, we distinguish between two possible explanations of the effect found by Nieuwland and van Berkum (2006): lexical priming alone, and event-level priming that may include lexical priming.

As discussed, language models can be used to model the latter. To model the former, we turn to word vectors—representations of words derived from their co-occurrence statistics, either directly or based on word embeddings learned by neural networks (see, e.g., Dumais, Furnas, Landauer, Deerwester, & Harshman, 1988; Landauer, Foltz, & Laham, 1998; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Pennington, Socher, & Manning, 2014; Mikolov, Grave, Bojanowski, Puhrsch, & Joulin, 2018; Tulkens, Emmery, & Daelemans, 2016; Grave, Bojanowski, Gupta, Joulin, & Mikolov, 2018). The cosine distance between the vector of each critical word (e.g. *in love* or *salted*) and the mean of the vectors of the words in the preceding context can therefore be used to test how similar the critical word is to the words preceding it (Ettinger et al., 2016; Uchida et al., 2021), and thereby model the effects of lexical priming alone. To do this this we used three sets of Dutch word vectors (from Tulkens et al., 2016; and Grave et al., 2018).

## Background

A number of researchers have attempted to model the N400 computationally, including using language models (Parviz et al., 2011; Frank et al., 2015; Aurnhammer & Frank, 2019b; Michaelov & Bergen, 2020; Merkx & Frank, 2021; Michaelov et al., 2021, 2022; Szewczyk & Federmeier, 2022; Michaelov et al., 2023) and the distances between vector representations of words (Parviz et al., 2011; Van Petten, 2014; Ettinger et al., 2016; Uchida et al., 2021; Michaelov et al., 2023). There have also been several attempts to computationally model whether the amplitude of the N400 response is impacted by situation models (Uchida et al., 2021; Venhuizen et al., 2019) and thematic roles (Brouwer, Crocker, Venhuizen, & Hoeks, 2017; Fitz & Chang, 2019; Rabovsky, Hansen, & McClelland, 2018).

To our knowledge, only one previous study (Uchida et al., 2021) has directly attempted to model the discourse effect found by Nieuwland and van Berkum (2006), and it does not rely on purely distributional linguistic information. Uchida et al. (2021) base their model on Wikipedia2Vec (Yamada et al., 2020) vectors—while these include distributional information derived from the surface-level statistics of language, they also include information about hyperlinks between Wikipedia pages, and thus structured semantic relations based on human judgements of relevance and importance (Yamada et al., 2020). Additionally, Uchida et al. (2021) only look at the English-translated version of the single stimulus item presented in (1), and thus, it is unclear whether the results generalize to all the stimuli in the original study. The current study overcomes these inferential limitations by using the original Dutch stimuli and by using neural language models and word vectors trained only on natural language input.

## The present study

We investigate the adequacy of distributional knowledge to explain the human N400 effect found by Nieuwland and van

Berkum (2006) using predictions of neural network language models and the distance between the word vectors of the critical words and their context. Specifically, we ask this question for two possible variants of the effect found by Nieuwland and van Berkum (2006).

Nieuwland and van Berkum (2006) presented experimental participants with short stories such as those in (1) including "canonical" sentences like *the peanut was **salted*** or "noncanonical" ones like *the peanut was **in love***. One approach to whether language models and humans show the same prediction patterns (taken by Uchida et al., 2021) is to compare the statistical metrics the critical words elicit in the context of the full story versus in isolation. Without preceding context, these sentences should produce values that match the canonicality of the sentence, but the difference should attenuate or reverse following the story context.

Thus, we ran a statistical analysis testing for an interaction between stimulus length (full story or only the last sentence) and canonicality (canonical or noncanonical). Such an interaction would reveal a context-dependent difference in the effect of canonicality on our statistical metrics; and thus would replicate in neural language models the effect found by Nieuwland and van Berkum (2006).

However, an interaction between stimulus length and canonicality in this direction could result from either a reversal or a decrease in the magnitude of the canonicality effect. Canonical stimuli might elicit lower surprisals or smaller cosine distances in both context conditions, but of different magnitudes. For this reason, we label the effect measured by an interaction (in the expected direction) a **reduction effect**.

Nieuwland and van Berkum (2006) did not employ the 2 x 2 design that would allow them to detect an interaction—they compared the N400 in context only, finding that canonical stimuli actually elicited larger N400 responses than noncanonical stimuli. To replicate this finding, we test whether the canonical full-length stimuli elicit higher surprisals or greater cosine distances than the noncanonical full-length stimuli, a **reversal effect**.

If either language models or word vectors can successfully model the reversal effect, this would suggest that distributional information is sufficient to explain the data reported by Nieuwland and van Berkum (2006). Thus, while situation models and extralinguistic information may be involved in the neurocognitive system underlying the N400, additional evidence is required to prove this. If neither can model either effect, this would undermine the claim that distributional information is sufficient to explain the effect found by Nieuwland and van Berkum (2006). Finally, if either language models or word vectors can successfully model the reduction effect but not the full reversal effect, this may support the idea that distributional information could be used as part of the neurocognitive system underlying the N400 response, but that it is not sufficient to yield the dynamic contextual sensitivity humans display. Situation models and other sources of information might explain the remainder.

# Method

## Materials

Stimuli were used from the original experiment, and are provided online[1] by the authors (Nieuwland & van Berkum, 2006). We compared the effect of experimental condition on the N400 and on neural network surprisal (as in Michaelov & Bergen, 2020) and the cosine similarity between the word vector of the critical word and the mean of the word vectors in its context (as in Ettinger et al., 2016).

The stimuli use 60 full-length story frames, each of which has either a canonical or noncanonical predicate, for 120 unique stories. As the aim is to model human online comprehension processes, the models only used the text before the critical words (e.g., *in love* or *salted*) to predict the critical words, so stories were truncated after the critical word. For the critical sentence stimuli, we isolated the last sentence of these truncated stories, including and up to the critical word in each story (e.g., *The peanut was in love*). This produced 240 stimuli, as shown in Table 1.

| Predicate Type | Stimulus Length | Count |
|---|---|---|
| Canonical | Full-length | 60 |
| Canonical | critical sentence | 60 |
| Noncanonical | Full-length | 60 |
| Noncanonical | critical sentence | 60 |

Table 1: Experimental stimuli derived from Nieuwland and van Berkum (2006).

## Statistical Analysis

Statistical analysis and data manipulation were carried out in *R* (R Core Team, 2020) using *Rstudio* (RStudio Team, 2020) and the *tidyverse* (Wickham et al., 2019) and *lme4* (Bates, Mächler, Bolker, & Walker, 2015) packages. Code, data, and statistical analyses are provided at `https://osf.io/wnj76`.

# Experiment 1: Language Models

## Language models

We used six pretrained models available through the *transformers* package (Wolf et al., 2020). These were all of the available monolingual Dutch language models using standard architectures and training procedures at the time of analysis. Four of these models—Dutch versions of the medium (Havinga, 2021) and large (Havinga, 2022a) GPT-2 models (Radford et al., 2019) and Dutch versions of the 125 million parameter (Havinga, 2022b) and 1.3 billion parameter (Havinga, 2022c) GPT-Neo models (Black, Gao, Wang, Leahy, & Biderman, 2021)—were autoregressive, meaning that they are trained to predict a word based only on its preceding context. The remaining two models—BERTje (de Vries et al., 2019) and RobBERT v2 (Delobelle et al., 2020),

[1] `https://www.researchgate.net/publication/268208198`

based on BERT (Devlin, Chang, Lee, & Toutanova, 2019) and RoBERTa (Liu et al., 2019), respectively—are masked language models, meaning that they are also trained to predict a word based on the text following the critical word. However, as stated, in the present study, all models were only provided with the context preceding the critical words. We ran the stimuli through each language model, calculating the surprisal of each critical word that was in the model's vocabulary (we restricted our analyses to these items). To do this, we calculated the negative of the logarithm of the probabilities provided for each critical word by each of the language models. We then tested for the reduction and reversal effects with these surprisal values. The language models were run in *Python* (Van Rossum & Drake, 2009), using the *PyTorch* (Paszke et al., 2019) implementation of each model, as provided by the *transformers* package (Wolf et al., 2020).

## Reduction effect

In order to test the reduction effect, we constructed linear mixed-effects regression models, with the surprisal calculated from each language model as the dependent variable. In each model, predicate type (canonical or noncanonical) and stimulus length (full-length or critical sentence) were fixed effects and story frame (each of the 60) was a random intercept. For the regressions with the autoregressive models and BERTje surprisal as their dependent variables, we then constructed regressions also including an interaction between predicate type and stimulus length. Using likelihood ratio tests, we found that these regressions including the interaction fit the data significantly better than those without the interaction (GPT-2 Medium: $\chi^2(1) = 112.0, p < 0.001$; GPT-2 Large: $\chi^2(1) = 115.9, p < 0.001$; GPT-Neo 125M: $\chi^2(1) = 67.3, p < 0.001$; GPT-Neo 1.3B: $\chi^2(1) = 56.3, p < 0.001$; BERTje: $\chi^2(1) = 44.4, p < 0.001$), indicating a significant interaction between predicate type and stimulus length. The regression with RobBERT surprisal as its dependent variable and no interaction had a singular fit, but the regression with the interaction did not. Thus, instead of running a likelihood ratio test to investigate whether there was a significant interaction, we used a a Type III ANOVA with Satterthwaite's method for estimating degrees of freedom (Kuznetsova, Brockhoff, & Christensen, 2017) on the regression with the interaction, finding it to be a significant predictor of RobBERT surprisal ($F(1, 71.2) = 81.1, p < 0.001$). Note that all reported $p$-values are corrected for multiple comparisons based on false discovery rate (Benjamini & Yekutieli, 2001).

For all language models, there was a significant interaction between predicate type and stimulus length. Further inspection of the regressions showed that in all cases, the interaction was in the expected direction. Thus, all models displayed the reduction effect. This can be seen visually in Figure 1—in all models, when only the critical sentence was presented, the mean surprisal for critical words in canonical sentences is lower than for critical words in noncanonical sentences. Conversely, when the full-length story is presented to the language models, the critical words in the noncanonical sentences elicit a lower or roughly-equal surprisal than the critical words in the canonical sentences.

## Reversal effect

To test for which models this latter finding was statistically significant, we initially attempted to fit linear mixed-effects regression models for each the full-length and critical sentence stimulus results for each language model; however, this led to several models with singular fits. Instead, we carried out pairwise two-tailed $t$-tests, comparing the surprisal of canonical and noncanonical stimuli for full-length and critical sentence stimuli for each language model.

First, we test whether the decontextualized canonical critical sentence stimuli elicit significantly lower surprisals than noncanonical critical sentence stimuli. After correction for multiple comparisons, they do so in all language models (GPT-2 Medium: $t(88.7) = -9.91$, $p < 0.001$; GPT-2 Large: $t(88.1) = -10.1$, $p < 0.001$; GPT-Neo 125M: $t(88.6) = -10.3$, $p < 0.001$ ; GPT-Neo 1.3B: $t(85.5) = -9.62$, $p < 0.001$; BERTje: $t(48.4) = -5.99$, $p < 0.001$; RobBERT: $t(55.1) = -7.67$, $p < 0.001$).

Next, in order to investigate the reversal effect, we test whether canonical full-length stimuli elicit lower surprisals than noncanonical full-length stimuli. After correction for multiple comparisons, only the Dutch GPT-2 models successfully model the reversal effect—they are the only models for which canonical full-length stimuli elicit significantly higher surprisals than noncanonical full-length stimuli (GPT-2 Medium: $t(86.3) = 6.11$, $p < 0.001$; GPT-2 Large: $t(88.4) = 5.65$, $p < 0.001$).

The difference in other models was not significant after correction for multiple comparisons (GPT-Neo 125M: $t(88.9) = -0.77$, $p = 1.000$ ; GPT-Neo 1.3B: $t(88.9) = 0.47$, $p = 1.000$; BERTje: $t(51.5) = 0.79$, $p = 1.000$; RobBERT: $t(46.6) = 2.32$, $p = 0.120$).

However, it is worth noting that the contrast between the two sets of results (critical sentence only vs. full stimulus) means that significant canonicality effects for the critical sentence stimuli disappear in the full-length stimuli, underscoring the presence of a reduction effect in the Dutch GPT-Neo models, BERTje, and RobBERT.

## Discussion

Nieuwland and van Berkum (2006) found that in a suitably supportive context, noncanonical stimuli like *de pinda was **verliefd*** ('the peanut was **in love**') elicit smaller N400 responses than canonical stimuli such as *de pinda was **gezouten*** ('the peanut was **salted**')—context not only mitigated but reversed the effect of animacy violation.

We find that two language models also display this reversal effect: Dutch GPT-2 Medium (Havinga, 2021) and Dutch GPT-2 Large (Havinga, 2022a). When these models are presented with the same contexts, the surprisal of critical words in the noncanonical condition is lower than that elicited by those in the canonical condition.

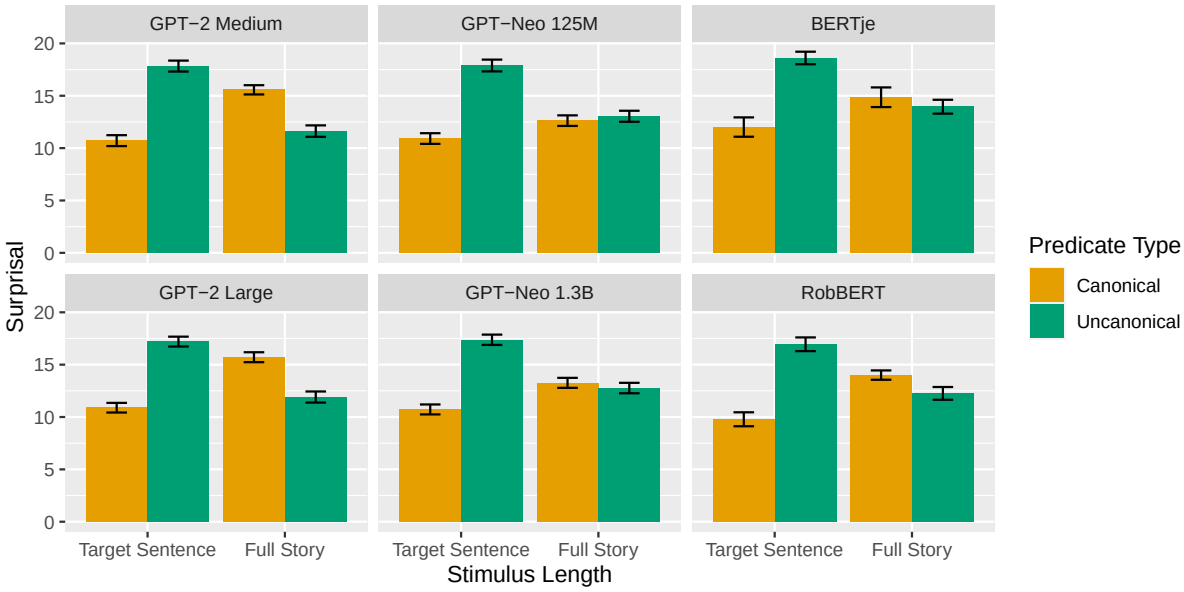


Figure 1: Surprisal elicited by critical words for each predicate type and stimulus length.

This is not the case for the remaining four language models: Dutch GPT-Neo 125M (Havinga, 2022b), Dutch GPT-Neo 1.3B (Havinga, 2022c), BERTje (de Vries et al., 2019), and RobBERT (Delobelle et al., 2020). However, these models do display the weaker reduction effect, and further, the absence of a significant difference between conditions for these models when presented with the full stories shows that the difference between canonical and noncanonical critical sentence stimuli is not just reduced, but disappears entirely.

It may be tempting to infer that the architecture of autoregressive transformers, and in particular, those based on the GPT-2 architecture, leads to success capturing the effect. However, it should be noted that before correction for multiple comparisons, RobBERT also successfully displays the reversal effect. In addition, not all language models had the same vocabulary, and thus, a different number of items were analyzed across models[2]. For these reasons, and because these models are all of various sizes and trained on several different datasets, we believe it would be premature to draw conclusions about how language model architecture impacts whether a model displays the reversal effect.

## Experiment 2: Word Vectors

### Cosine Distance

In this study, we used 3 sets of pretrained word vectors: the 300-dimensional Dutch *fastText* vectors (Grave et al., 2018) trained on Dutch text from Wikipedia[3] and Common Crawl[4] and two 320-dimensional Dutch word vectors released by Tulkens et al. (2016)—one trained on *COW* (COrpora from the Web; Schäfer & Bildhauer, 2012) and one trained on a *Combined* corpus made up of the SoNaR corpus (Oostdijk, Reynaert, Hoste, & Schuurman, 2013) and text from Wikipedia and Roularta[5]. Cosine distance was calculated (using *SciPy*; Virtanen et al., 2020) between the mean of the word vectors for all words in the preceding context and the word embedding for the critical word. All critical words were present in the vectors, so all experimental items were included in the analysis; it should be noted though that words in the context that were not present in the vectors were ignored when calculating cosine distance. The cosine distances for critical words in each condition are shown in Figure 2.

### Reduction effect

As with language model surprisal, we constructed linear mixed-effects regressions with predicate type and stimulus length as fixed effects and story from as a random intercept. With these models, the cosine distance calculated using each set of word vectors was the dependent variable. The interaction between predicate type and stimulus length was significant for all vectors after correcting for multiple comparisons (fastText: $\chi^2(1) = 12.0, p = 0.003$; Combined: $\chi^2(1) = 40.8, p < 0.001$; COW: $\chi^2(1) = 66.4, p < 0.001$).

### Reversal effect

When comparing the cosine distances calculated between the embedding of the critical words and the preceding words of the critical sentence using two-tailed *t*-tests as with surprisal, there was a significant difference between canonical

[2]Though it should be noted that an alternate analysis including all critical words by operationalizing the suprisal of multi-token words as the sum of their tokens' surprisals (see Michaelov & Bergen, 2022b) shows the same qualitative results for all models except for BERTje—which performs worse.

[3]https://nl.wikipedia.org/

[4]https://commoncrawl.org/

[5]https://www.roularta.be

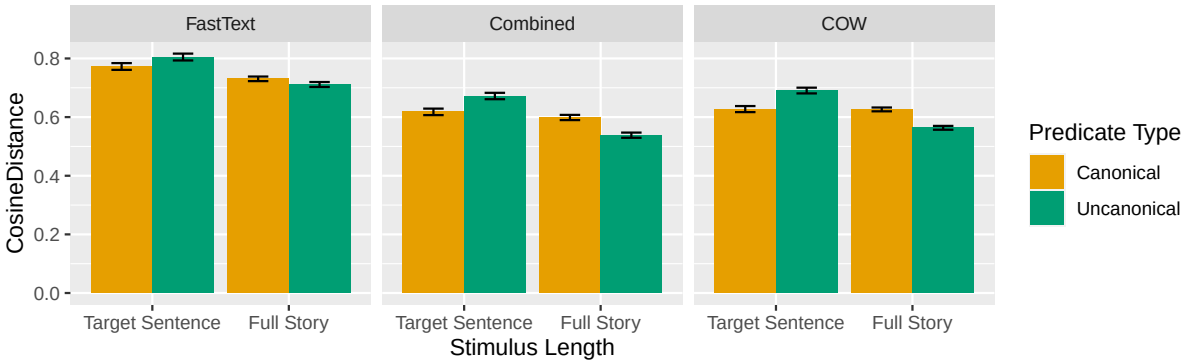


Figure 2: Cosine distance elicited by critical words for each predicate type and stimulus length.

and noncanonical critical words for Combined and COW vectors (Combined: $t(116.9) = -3.48$, $p = 0.004$; COW: $t(116.5) = -4.45$, $p < 0.001$), but not fastText vectors (fastText: $t(118.0) = -1.96$, $p = 0.237$).

Similarly, when comparing the cosine distances between the critical word and the preceding words of the full story, there was a significant difference between canonical and noncanonical critical words for Combined and COW vectors (Combined: $t(117.0) = 4.82$, $p < 0.001$; COW: $t(117.0) = 6.78$, $p < 0.001$), but not fastText vectors (fastText: $t(117.4) = 1.68$, $p = 0.418$).

### Discussion

The cosine distances calculated from all three sets of word vectors displayed the reduction effect, and two out of three displayed the reversal effect. Thus, the results suggest that the N400 effect reported by Nieuwland and van Berkum (2006) can be explained by lexical priming based on distributional linguistic knowledge alone.

The present study corroborates the finding of Uchida et al. (2021), and expands upon it in several ways. First, we explicitly tested for the reversal effect—not just whether canonical and noncanonical stimuli differ depending on whether there is a preceding story or not, but also whether the noncanonical sentence is more expected than the canonical when the story is present. Second, we found that word vector cosine distance can model the effect for multiple stimuli, not just the *peanut was in love* example. Third, we found that the effect can be modeled in Dutch, the language in which the human study was carried out. And finally, we found that vectors derived from text data only (i.e., without additional information) are able to model the effect.

## General Discussion

Human comprehenders use context to update expectations about upcoming words, making a sentence that would be highly unlikely on its own more predictable than a sentence that would be relatively likely on its own. More strikingly, humans do this even when the event described is implausible, violating the constraint, for instance, that only animate, conscious entities can fall in love. The human comprehension system is quite flexible if it can update expectations about what peanuts, for example, can do, based only a story that indirectly implies the animacy of a fictional peanut.

It has often been assumed that this flexibility requires situation models that are explicitly structured (Venhuizen et al., 2019) or involve non-linguistic world knowledge (Uchida et al., 2021). However, the present findings show that it is possible for purely linguistic language models model with no direct experiential grounding to update their expectations based on the linguistic context and knowledge of the statistics of language. Thus, the dynamics of event-level priming based on the distributional statistics of language may in some implicit, unspecified way approximate the effects on language comprehension previously ascribed to situation models.

In fact, the results of the present study provide evidence for an even simpler explanation. Within final sentences alone, canonical critical words were more similar to their contexts than noncanonical words, but when we include the full story context, it is the noncanonical critical words that are more similar to their contexts. It is already well-established that the amplitude of the N400 to a given word is reduced when it is semantically related to a previously-seen word (Bentin, McCarthy, & Wood, 1985; Rugg, 1985; Van Petten & Kutas, 1988; Kutas & Hillyard, 1989; Holcomb, 1988; Kutas, 1993; Lau, Holcomb, & Kuperberg, 2013). Overall, then, our results show that in principle, it is possible that the pattern in the N400 responses reported by Nieuwland and van Berkum (2006) may not rely on situation models or even event-level priming, but rather reflect some form of lexical priming.

It may still be the case that humans use structured or semantically-rich situation models in online language comprehension (see, e.g., Kuperberg et al., 2020). However, the results of the study carried out by Nieuwland and van Berkum (2006) appear to provide weaker evidence for this than previously believed. Language model predictions or even lexical priming based on language statistics appear to be sufficient to explain the effect, at least qualitatively—a valuable line of future research would be to test whether these can fully account for the effect in single-trial N400 data.

## Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments. This work was partially supported by a 2021-2022 Center for Academic Research and Training in Anthropogeny Annette Merle-Smith Fellowship awarded to James A. Michaelov.

## References

Aravena, P., Courson, M., Frak, V., Cheylus, A., Paulignan, Y., Deprez, V., & Nazir, T. (2014). Action relevance in linguistic context drives word-induced motor activity. *Frontiers in Human Neuroscience*, *8*. Retrieved 2022-08-19, from https://www.frontiersin.org/articles/10.3389/fnhum.2014.00163

Aurnhammer, C., & Frank, S. L. (2019a). Comparing Gated and Simple Recurrent Neural Network Architectures as Models of Human Sentence Processing. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society (CogSci 2019)*. Retrieved 2022-06-02, from https://cogsci.mindmodeling.org/2019/papers/0041/ doi: 10.31234/osf.io/wec74

Aurnhammer, C., & Frank, S. L. (2019b). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, *134*, 107198. Retrieved 2020-09-07, from http://www.sciencedirect.com/science/article/pii/S0028393219302404 doi: 10.1016/j.neuropsychologia.2019.107198

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Benjamini, Y., & Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics*, *29*(4), 1165–1188. Retrieved 2021-05-04, from https://www.jstor.org/stable/2674075

Bentin, S., McCarthy, G., & Wood, C. C. (1985). Event-related potentials, lexical decision and semantic priming. *Electroencephalography and Clinical Neurophysiology*, *60*(4), 343–355. Retrieved 2022-01-19, from https://www.sciencedirect.com/science/article/pii/0013469485900082 doi: 10.1016/0013-4694(85)90008-2

Bicknell, K., Elman, J. L., Hare, M., McRae, K., & Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language*, *63*(4), 489–505. Retrieved 2020-06-26, from http://www.sciencedirect.com/science/article/pii/S0749596X10000653 doi: 10.1016/j.jml.2010.08.004

Black, S., Gao, L., Wang, P., Leahy, C., & Biderman, S. (2021). *GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.5297715 doi: 10.5281/zenodo.5297715

Bransford, J. D., Barclay, J. R., & Franks, J. J. (1972). Sentence memory: A constructive versus interpretive approach. *Cognitive Psychology*, *3*(2), 193–209. Retrieved 2021-11-09, from https://www.sciencedirect.com/science/article/pii/0010028572900035 doi: 10.1016/0010-0285(72)90003-5

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. J. (2017). A Neurocomputational Model of the N400 and the P600 in Language Processing. *Cognitive Science*, *41*(S6), 1318–1352. Retrieved 2020-02-24, from https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12461 doi: 10.1111/cogs.12461

Butcher, K. R., & Kintsch, W. (2012). Text Comprehension and Discourse Processing. In *Handbook of Psychology, Second Edition* (chap. 21). American Cancer Society. Retrieved 2021-11-09, from https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118133880.hop204021 doi: 10.1002/9781118133880.hop204021

Chwilla, D. J., & Kolk, H. H. J. (2005). Accessing world knowledge: Evidence from N400 and reaction time priming. *Cognitive Brain Research*, *25*(3), 589–606. Retrieved 2020-06-25, from http://www.sciencedirect.com/science/article/pii/S0926641005002259 doi: 10.1016/j.cogbrainres.2005.08.011

de Groot, A. M. B. (2011). *Language and cognition in bilinguals and multilinguals an introduction*. New York [u.a.: Psychology Press.

de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., & Nissim, M. (2019). BERTje: A Dutch BERT Model. *arXiv:1912.09582 [cs]*. Retrieved 2021-05-27, from http://arxiv.org/abs/1912.09582

Delobelle, P., Winters, T., & Berendt, B. (2020). RobBERT: A Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 3255–3265). Online: Association for Computational Linguistics. Retrieved 2021-05-27, from https://www.aclweb.org/anthology/2020.findings-emnlp.292 doi: 10.18653/v1/2020.findings-emnlp.292

DeLong, K. A., Troyer, M., & Kutas, M. (2014). Pre-Processing in Sentence Comprehension: Sensitivity to Likely Upcoming Meaning and Structure. *Language and Linguistics Compass*, *8*(12), 631–645. Retrieved 2020-06-25, from https://onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12093 doi: 10.1111/lnc3.12093

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (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)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computa-

tional Linguistics. Retrieved 2020-11-24, from https://www.aclweb.org/anthology/N19-1423 doi: 10.18653/v1/N19-1423

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '88* (pp. 281–285). Washington, D.C., United States: ACM Press. Retrieved 2021-02-01, from http://portal.acm.org/citation.cfm?doid=57167.57214 doi: 10.1145/57167.57214

Ettinger, A., Feldman, N., Resnik, P., & Phillips, C. (2016). Modeling N400 amplitude using vector space models of word representation. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society.* Philadelphia, USA. Retrieved from https://cogsci.mindmodeling.org/2016/papers/0256/

Ferguson, H. J., & Sanford, A. J. (2008). Anomalies in real and counterfactual worlds: An eye-movement investigation. *Journal of Memory and Language*, *58*(3), 609–626. Retrieved 2022-08-19, from https://www.sciencedirect.com/science/article/pii/S0749596X07000770 doi: 10.1016/j.jml.2007.06.007

Ferguson, H. J., Sanford, A. J., & Leuthold, H. (2008). Eye-movements and ERPs reveal the time course of processing negation and remitting counterfactual worlds. *Brain Research*, *1236*, 113–125. Retrieved 2022-08-19, from https://www.sciencedirect.com/science/article/pii/S0006899308018295 doi: 10.1016/j.brainres.2008.07.099

Filik, R., & Leuthold, H. (2008). Processing local pragmatic anomalies in fictional contexts: Evidence from the N400. *Psychophysiology*, *45*(4), 554–558. Retrieved 2022-08-19, from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8986.2008.00656.x doi: 10.1111/j.1469-8986.2008.00656.x

Fischler, I., & Bloom, P. A. (1979). Automatic and attentional processes in the effects of sentence contexts on word recognition. *Journal of Verbal Learning and Verbal Behavior*, *18*(1), 1–20. Retrieved 2021-06-03, from https://www.sciencedirect.com/science/article/pii/S0022537179905346 doi: 10.1016/S0022-5371(79)90534-6

Fitz, H., & Chang, F. (2019). Language ERPs reflect learning through prediction error propagation. *Cognitive Psychology*, *111*, 15–52. Retrieved 2019-12-04, from https://linkinghub.elsevier.com/retrieve/pii/S0010028518300124 doi: 10.1016/j.cogpsych.2019.03.002

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11. Retrieved 2019-12-04, from http://www.sciencedirect.com/science/article/pii/S0093934X14001515 doi: 10.1016/j.bandl.2014.10.006

Garnham, A. (1981). Mental models as representations of text. *Memory & Cognition*, *9*(6), 560–565. Retrieved 2021-11-04, from https://doi.org/10.3758/BF03202350 doi: 10.3758/BF03202350

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning Word Vectors for 157 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* Miyazaki, Japan: European Language Resources Association (ELRA). Retrieved 2020-09-28, from https://www.aclweb.org/anthology/L18-1550

Hagoort, P., & van Berkum, J. (2007). Beyond the sentence given. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 801–811. Retrieved 2022-08-19, from https://royalsocietypublishing.org/doi/10.1098/rstb.2007.2089 doi: 10.1098/rstb.2007.2089

Havinga, Y. (2021). *GPT2-Medium pre-trained on cleaned Dutch mC4.* Retrieved 2022-01-26, from https://huggingface.co/yhavinga/gpt2-medium-dutch

Havinga, Y. (2022a). *GPT2-Large pre-trained on cleaned Dutch mC4.* Retrieved 2022-01-26, from https://huggingface.co/yhavinga/gpt2-large-dutch

Havinga, Y. (2022b). *GPT-Neo 125M pre-trained on cleaned Dutch mC4.* Retrieved 2022-01-26, from https://huggingface.co/yhavinga/gpt-neo-125M-dutch

Havinga, Y. (2022c). *GPT Neo 1.3B pre-trained on cleaned Dutch mC4.* Retrieved 2022-01-26, from https://huggingface.co/yhavinga/gpt-neo-1.3B-dutch

Hoeben Mannaert, L., & Dijkstra, K. (2021). Situation model updating in young and older adults. *International Journal of Behavioral Development*, *45*(5), 389–396. Retrieved 2021-11-08, from https://doi.org/10.1177/0165025419874125 doi: 10.1177/0165025419874125

Holcomb, P. J. (1988). Automatic and attentional processing: An event-related brain potential analysis of semantic priming. *Brain and Language*, *35*(1), 66–85. Retrieved 2022-01-19, from https://www.sciencedirect.com/science/article/pii/0093934X88901010 doi: 10.1016/0093-934X(88)90101-0

Johnson-Laird, P. N. (1980). Mental models in cognitive science. *Cognitive Science*, *4*(1), 71–115. Retrieved 2021-11-04, from https://www.sciencedirect.com/science/article/pii/S0364021381800055 doi: 10.1016/S0364-0213(81)80005-5

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness* (No. 6). Harvard University Press.

Kassner, N., & Schütze, H. (2020). Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational*

*Linguistics* (pp. 7811–7818). Online: Association for Computational Linguistics. Retrieved 2022-05-10, from https://aclanthology.org/2020.acl-main.698 doi: 10.18653/v1/2020.acl-main.698

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163. Retrieved 2021-11-09, from https://psycnet.apa.org/fulltext/1988-28529-001.pdf doi: 10.1037/0033-295X.95.2.163

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York, NY, US: Cambridge University Press.

Kintsch, W. (2005). An Overview of Top-Down and Bottom-Up Effects in Comprehension: The CI Perspective. *Discourse Processes*, *39*(2-3), 125–128. Retrieved 2021-11-04, from https://doi.org/10.1080/0163853X.2005.9651676 doi: 10.1080/0163853X.2005.9651676

Kintsch, W. (2018). Revisiting the Construction—Integration Model of Text Comprehension and its Implications for Instruction. In *Theoretical Models and Processes of Literacy* (Seventh ed.). Routledge.

Kintsch, W., & Mangalath, P. (2011). The Construction of Meaning. *Topics in Cognitive Science*, *3*(2), 346–370. Retrieved 2021-11-04, from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1756-8765.2010.01107.x doi: 10.1111/j.1756-8765.2010.01107.x

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363–394. doi: 10.1037/0033-295X.85.5.363

Kuperberg, G. R., Brothers, T., & Wlotko, E. W. (2020). A Tale of Two Positivities and the N400: Distinct Neural Signatures Are Evoked by Confirmed and Violated Predictions at Different Levels of Representation. *Journal of Cognitive Neuroscience*, *32*(1), 12–35. Retrieved 2020-06-26, from https://www.mitpressjournals.org/doi/abs/10.1162/jocn_a_01465 doi: 10.1162/jocn_a_01465

Kutas, M. (1993). In the company of other words: Electrophysiological evidence for single-word and sentence context effects. *Language and Cognitive Processes*, *8*(4), 533–572. Retrieved 2020-04-07, from https://doi.org/10.1080/01690969308407587 doi: 10.1080/01690969308407587

Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. In M. Bar (Ed.), *Predictions in the brain: Using our past to generate a future* (pp. 190–207). New York, NY, US: Oxford University Press. doi: 10.1093/acprof:oso/9780195395518.003.0065

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163. Retrieved 2020-01-31, from http://www.nature.com/articles/307161a0 doi: 10.1038/307161a0

Kutas, M., & Hillyard, S. A. (1989). An Electrophysiological Probe of Incidental Semantic Association. *Journal of Cognitive Neuroscience*, *1*(1), 38–49. Retrieved 2022-05-23, from https://doi.org/10.1162/jocn.1989.1.1.38 doi: 10.1162/jocn.1989.1.1.38

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*, 1–26. Retrieved 2022-02-03, from https://doi.org/10.18637/jss.v082.i13 doi: 10.18637/jss.v082.i13

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2-3), 259–284. Retrieved 2021-02-01, from http://www.tandfonline.com/doi/abs/10.1080/01638539809545028 doi: 10.1080/01638539809545028

Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 Effects of Prediction from Association in Single-word Contexts. *Journal of Cognitive Neuroscience*, *25*(3), 484–502. Retrieved 2021-01-23, from https://www.mitpressjournals.org/doi/abs/10.1162/jocn_a_00328 doi: 10.1162/jocn_a_00328

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. Retrieved 2020-05-28, from http://www.sciencedirect.com/science/article/pii/S0010027707001436 doi: 10.1016/j.cognition.2007.05.006

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*. Retrieved 2020-11-24, from http://arxiv.org/abs/1907.11692

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22–60. Retrieved 2020-05-29, from http://www.sciencedirect.com/science/article/pii/S0010028516301384 doi: 10.1016/j.cogpsych.2016.06.002

Menenti, L., Petersson, K. M., Scheeringa, R., & Hagoort, P. (2009). When Elephants Fly: Differential Sensitivity of Right and Left Inferior Frontal Gyri to Discourse and World Knowledge. *Journal of Cognitive Neuroscience*, *21*(12), 2358–2368. Retrieved 2022-08-19, from https://doi.org/10.1162/jocn.2008.21163 doi: 10.1162/jocn.2008.21163

Merkx, D., & Frank, S. L. (2021). Human Sentence Processing: Recurrence or Attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 12–22). Online: Association for Computational Linguistics. Retrieved 2021-07-15, from https://aclanthology.org/2021.cmcl-1.2 doi: 10.18653/v1/2021.cmcl-1.2

Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., & Elman, J. L. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language*, *66*(4), 545–567. Retrieved 2020-06-25, from http://www.sciencedirect

.com/science/article/pii/S0749596X12000034 doi: 10.1016/j.jml.2012.01.001

Michaelov, J. A., Bardolph, M. D., Coulson, S., & Bergen, B. K. (2021). Different kinds of cognitive plausibility: Why are transformers better than RNNs at predicting N400 amplitude? In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society* (pp. 300–306). University of Vienna, Vienna, Austria (Hybrid).

Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., & Coulson, S. (2023). Strong Prediction: Language model surprisal explains multiple N400 effects. *Neurobiology of Language*, 1–71. Retrieved 2023-04-25, from https://doi.org/10.1162/nol_a_00105 doi: 10.1162/nol_a_00105

Michaelov, J. A., & Bergen, B. (2022a). Collateral facilitation in humans and language models. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)* (pp. 13–26). Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics. Retrieved 2023-02-07, from https://aclanthology.org/2022.conll-1.2

Michaelov, J. A., & Bergen, B. K. (2020). How well does surprisal explain N400 amplitude under different experimental conditions? In *Proceedings of the 24th Conference on Computational Natural Language Learning* (pp. 652–663). Online: Association for Computational Linguistics. Retrieved 2021-02-02, from https://www.aclweb.org/anthology/2020.conll-1.53 doi: 10.18653/v1/2020.conll-1.53

Michaelov, J. A., & Bergen, B. K. (2022b). The more human-like the language model, the more surprisal is the best predictor of N400 amplitude. In *NeurIPS 2022 Workshop on Information-Theoretic Principles in Cognitive Systems*. Retrieved 2022-12-16, from https://openreview.net/forum?id=uCgYvb8GNQZ

Michaelov, J. A., Coulson, S., & Bergen, B. K. (2022). So Cloze yet so Far: N400 Amplitude is Better Predicted by Distributional Information than Human Predictability Judgements. *IEEE Transactions on Cognitive and Developmental Systems*. doi: 10.1109/TCDS.2022.3176783

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2018). Advances in Pre-Training Distributed Word Representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). Retrieved 2022-05-26, from https://aclanthology.org/L18-1008

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26* (pp. 3111–3119). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

Misra, K., Ettinger, A., & Rayz, J. (2020). Exploring BERT's Sensitivity to Lexical Cues using Tests from Semantic Priming. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 4625–4635). Online: Association for Computational Linguistics. Retrieved 2021-01-26, from https://www.aclweb.org/anthology/2020.findings-emnlp.415 doi: 10.18653/v1/2020.findings-emnlp.415

Nieuwland, M. S., & van Berkum, J. J. A. (2006). When Peanuts Fall in Love: N400 Evidence for the Power of Discourse. *Journal of Cognitive Neuroscience*, *18*(7), 1098–1111. Retrieved 2020-02-02, from https://doi.org/10.1162/jocn.2006.18.7.1098 doi: 10.1162/jocn.2006.18.7.1098

Oostdijk, N., Reynaert, M., Hoste, V., & Schuurman, I. (2013). The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch. In P. Spyns & J. Odijk (Eds.), *Essential Speech and Language Technology for Dutch: Results by the STEVIN programme* (pp. 219–247). Berlin, Heidelberg: Springer. Retrieved 2023-01-12, from https://doi.org/10.1007/978-3-642-30910-6_13 doi: 10.1007/978-3-642-30910-6_13

Parviz, M., Johnson, M., Johnson, B., & Brock, J. (2011). Using Language Models and Latent Semantic Analysis to Characterise the N400m Neural Response. In *Proceedings of the Australasian Language Technology Association Workshop 2011* (pp. 38–46). Canberra, Australia. Retrieved 2020-10-06, from https://www.aclweb.org/anthology/U11-1007

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems* (Vol. 32). Curran Associates, Inc. Retrieved 2021-08-18, from https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics. Retrieved 2022-05-26, from https://aclanthology.org/D14-1162 doi: 10.3115/v1/D14-1162

R Core Team. (2020). *R: A language and environment for statistical computing* [Manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, *2*(9), 693–705. Retrieved 2019-12-04, from https://www.nature.com/articles/s41562-018-0406-4 doi: 10.1038/s41562-018-0406-4

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. , 24.

Radvansky, G. A., Zwaan, R. A., Federico, T., & Franklin, N. (1998). Retrieval from temporally organized situation models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(5), 1224–1237. doi: 10.1037/0278-7393.24.5.1224

Rosenbach, A. (2008). Animacy and grammatical variation—Findings from English genitive variation. *Lingua*, *118*(2), 151–171. Retrieved 2022-08-19, from https://www.sciencedirect.com/science/article/pii/S002438410700023X doi: 10.1016/j.lingua.2007.02.002

RStudio Team. (2020). *RStudio: Integrated development environment for r* [Manual]. Boston, MA. Retrieved from http://www.rstudio.com/

Rugg, M. D. (1985). The Effects of Semantic Priming and Word Repetition on Event-Related Potentials. *Psychophysiology*, *22*(6), 642–647. Retrieved 2022-01-19, from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8986.1985.tb01661.x doi: 10.1111/j.1469-8986.1985.tb01661.x

Schäfer, R., & Bildhauer, F. (2012). Building Large Corpora from the Web Using a New Efficient Tool Chain. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 486–493). Istanbul, Turkey: European Language Resources Association (ELRA). Retrieved 2023-01-12, from http://www.lrec-conf.org/proceedings/lrec2012/pdf/834_Paper.pdf

Szewczyk, J. M., & Federmeier, K. D. (2022). Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability. *Journal of Memory and Language*, *123*, 104311. Retrieved 2022-01-03, from https://www.sciencedirect.com/science/article/pii/S0749596X21000942 doi: 10.1016/j.jml.2021.104311

Tulkens, S., Emmery, C., & Daelemans, W. (2016). Evaluating Unsupervised Dutch Word Embeddings as a Linguistic Resource. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 4130–4136). Portorož, Slovenia: European Language Resources Association (ELRA). Retrieved 2023-01-12, from https://aclanthology.org/L16-1652

Uchida, T., Lair, N., Ishiguro, H., & Dominey, P. F. (2021). A Model of Online Temporal-Spatial Integration for Immediacy and Overrule in Discourse Comprehension. *Neurobiology of Language*, *2*(1), 83–105. Retrieved 2022-01-03, from https://doi.org/10.1162/nol_a_00026 doi: 10.1162/nol_a_00026

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

Van Berkum, J. J. A., Koornneef, A. W., Otten, M., & Nieuwland, M. S. (2007). Establishing reference in language comprehension: An electrophysiological perspective. *Brain Research*, *1146*, 158–171. Retrieved 2021-11-04, from https://www.sciencedirect.com/science/article/pii/S0006899306019998 doi: 10.1016/j.brainres.2006.06.091

Van Petten, C. (2014). Examining the N400 semantic context effect item-by-item: Relationship to corpus-based measures of word co-occurrence. *International Journal of Psychophysiology*, *94*(3), 407–419. Retrieved 2020-12-18, from https://linkinghub.elsevier.com/retrieve/pii/S0167876014016377 doi: 10.1016/j.ijpsycho.2014.10.012

Van Petten, C., & Kutas, M. (1988). Tracking the time course of meaning activation. In S. Small, G. Cottrell, & M. Tanenhaus (Eds.), *Lexical Ambiguity Resolution in the Comprehension of Human Language* (pp. 431–475). San Mateo, California: Morgan Kaufmann Publishers. Retrieved 2022-05-24, from http://kutaslab.ucsd.edu/people/kutas/pdfs/1988.LAR.431.pdf

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176–190. Retrieved 2020-05-29, from http://www.sciencedirect.com/science/article/pii/S0167876011002819 doi: 10.1016/j.ijpsycho.2011.09.015

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. Scotts Valley, CA: CreateSpace.

Venhuizen, N. J., Crocker, M. W., & Brouwer, H. (2019). Expectation-based Comprehension: Modeling the Interaction of World Knowledge and Linguistic Experience. *Discourse Processes*, *56*(3), 229–255. Retrieved 2019-12-04, from https://doi.org/10.1080/0163853X.2018.1448677 doi: 10.1080/0163853X.2018.1448677

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., … SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, *17*, 261–272. doi: 10.1038/s41592-019-0686-2

Warren, T., McConnell, K., & Rayner, K. (2008). Effects of context on eye movements when reading about possible and impossible events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 1001–1010. Retrieved 2022-08-19, from http://doi.apa.org/getdoi.cfm?doi=10.1037/0278-7393.34.4.1001 doi: 10.1037/0278-7393.34.4.1001

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi: 10.21105/joss.01686

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., … Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–

45). Online: Association for Computational Linguistics. Retrieved 2021-05-27, from https://www.aclweb.org/anthology/2020.emnlp-demos.6 doi: 10.18653/v1/2020.emnlp-demos.6

Xiang, M., & Kuperberg, G. (2015). Reversing expectations during discourse comprehension. *Language, Cognition and Neuroscience*, *30*(6), 648–672. Retrieved 2021-10-13, from https://doi.org/10.1080/23273798.2014.995679 doi: 10.1080/23273798.2014.995679

Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y., & Matsumoto, Y. (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* (pp. 23–30). Online: Association for Computational Linguistics. Retrieved 2022-01-14, from https://aclanthology.org/2020.emnlp-demos.4 doi: 10.18653/v1/2020.emnlp-demos.4

Zacks, J. M., & Ferstl, E. C. (2016). Discourse Comprehension. In G. Hickok & S. L. Small (Eds.), *Neurobiology of Language* (pp. 661–673). San Diego: Academic Press. Retrieved 2021-11-04, from https://www.sciencedirect.com/science/article/pii/B9780124077942000535 doi: 10.1016/B978-0-12-407794-2.00053-5

Zwaan, R. A. (2014). Embodiment and language comprehension: Reframing the discussion. *Trends in Cognitive Sciences*, *18*(5), 229–234. Retrieved 2020-07-08, from https://linkinghub.elsevier.com/retrieve/pii/S1364661314000527 doi: 10.1016/j.tics.2014.02.008

Zwaan, R. A. (2016). Situation models, mental simulations, and abstract concepts in discourse comprehension. *Psychonomic Bulletin & Review*, *23*(4), 1028–1034. Retrieved 2021-11-04, from https://doi.org/10.3758/s13423-015-0864-x doi: 10.3758/s13423-015-0864-x

Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The Construction of Situation Models in Narrative Comprehension: An Event-Indexing Model. *Psychological Science*, *6*(5), 292–297. Retrieved 2021-11-09, from https://doi.org/10.1111/j.1467-9280.1995.tb00513.x doi: 10.1111/j.1467-9280.1995.tb00513.x

Zwaan, R. A., & Madden, C. J. (2004). Updating situation models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 283–288. doi: 10.1037/0278-7393.30.1.283

Zwaan, R. A., Magliano, J. P., & Graesser, A. C. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of experimental psychology: Learning, memory, and cognition*, *21*(2), 386.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162–185. doi: 10.1037/0033-2909.123.2