How well does surprisal explain N400 amplitude under different experimental conditions?

James A. Michaelov
Deparmtment of Cognitive Science
University of California, San Diego
j1michae@ucsd.edu

Benjamin K. Bergen
Deparmtment of Cognitive Science
University of California, San Diego
bkbergen@ucsd.edu

Abstract

We investigate the extent to which word surprisal can be used to predict a neural measure of human language processing difficulty—the N400. To do this, we use recurrent neural networks to calculate the surprisal of stimuli from previously published neurolinguistic studies of the N400. We find that surprisal can predict N400 amplitude in a wide range of cases, and the cases where it cannot do so provide valuable insight into the neurocognitive processes underlying the response.

1 Introduction

The N400 component of the event-related brain potential is generally understood to be a neural signal of processing difficulty (Kutas and Federmeier, 2011). After over 1,000 articles published on the topic, we know that all else being equal, an upcoming word that is supported by the semantics of the context will elicit a lower-amplitude N400 than a word that is not (Kutas and Federmeier, 2011; Kuperberg et al., 2020). However, despite the great amount of experimental research on the topic, many aspects of the N400 are still not well understood.

In addition to 'long-standing and recent linguistic [...] inputs' (Kutas and Federmeier, 2011, p. 641), the context that impacts N400 amplitude is thought to include factors such as world experience, attentional state, and mood (Kutas and Federmeier, 2011). Over the last decade, there have been a number of attempts to use computational modeling to test hypotheses about the neurocognitive processes underlying the N400 and how the aforementioned factors may impact its amplitude (Parviz et al., 2011; Laszlo and Plaut, 2012; Laszlo and Armstrong, 2014; Rabovsky and McRae, 2014; Frank et al., 2015; Ettinger et al., 2016; Cheyette and Plaut, 2017; Brouwer et al., 2017; Delaney-Busch et al., 2017; Rabovsky et al., 2018; Venhuizen et al., 2018; Fitz and Chang, 2019).

As the majority of experimental research on the N400 involves manipulating the relationship between the stimulus and the preceding linguistic context (Kutas and Federmeier, 2011), a computational account of how linguistic inputs impact N400 amplitude is a logical starting point. Language models are inherently models of linguistic prediction based only on language input. Since N400 amplitude reflects how unexpected an upcoming word is based on context, the predictions of a language model can be used to model how expected a word is based on the linguistic input, and thereby investigate the extent to which N400 amplitude is explainable by linguistic input alone.

Recent research has shown that *surprisal*, a measure of how unlikely a language model predicts the next word in sequence to be, correlates overall with N400 amplitude (Frank et al., 2015; Aurnhammer and Frank, 2019). Thus, to investigate the extent to which N400 amplitude is explained by linguistic input alone, we ask to what extent surprisal can explain the variance observed in N400 amplitude.

In order to investigate this, we run experimental stimuli from eleven experiments from six papers (Urbach and Kutas, 2010; Kutas, 1993; Ito et al., 2016; Osterhout and Mobley, 1995; Ainsworth-Darnell et al., 1998; Kim and Osterhout, 2005) through two recurrent neural network language models (Jozefowicz et al., 2016; Gulordava et al., 2018), systematically comparing the significant predictors of N400 amplitude and surprisal. We find that in the majority of cases, significant differences in surprisal predict significant differences in N400 amplitude, and discuss the implications of the cases where it does not.

2 Background

2.1 The N400

The N400 is a negative deflection in the event-related brain potential (ERP) that peaks roughly

400ms after the presentation of a stimulus (Kutas and Hillyard, 1980; Kutas and Federmeier, 2011). Most current accounts agree that N400 amplitude reflects processing difficulty for a specific lexical item, where a lower amplitude reflects prior activation of some of the semantic content associated with the word (Kutas and Federmeier, 2011; Kuperberg, 2016; Kuperberg et al., 2020).

Recent research has found that N400 amplitude '*decreases* with supportive context, but does not *increase* when predictions are violated' (DeLong and Kutas, 2020, p. 2, emphasis in original; see Kutas and Federmeier, 2011; Van Petten and Luka, 2012; Luke and Christianson, 2016; Kuperberg et al., 2020, for discussion). Crucially, therefore, we should not think of N400 amplitude as a general measure of prediction error. It is not the case that the N400 elicited by a word increases when the word is more semantically anomalous or unexpected based on the preceding context; rather, it is the case that N400 amplitude is reduced when the word is semantically congruous or predictable because it is facilitated by the preceding context.

This facilitation can occur in a large number of ways. All else being equal, words that are more semantically congruous, typical, or plausible completions of a sentence elicit lower N400 amplitudes than words that are more semantically incongruous, atypical, and implausible completions, respectively (e.g. Kutas and Hillyard, 1980; Urbach and Kutas, 2010; Ito et al., 2016; Osterhout and Mobley, 1995; Ainsworth-Darnell et al., 1998; Kim and Osterhout, 2005; Kutas and Federmeier, 2011).

One well-known correlate of N400 amplitude is the cloze probability (Taylor, 1953; Bloom and Fischler, 1980) of a word—the probability that it will be offered to fill a specific gap in a sentence by a given sample of individuals in a norming study. All else being equal, higher-cloze completions elicit lower N400 amplitudes (Kutas and Hillyard, 1984; Kutas and Federmeier, 2011). Additionally, even when matched for cloze, words semantically related to the highest-cloze completion elicit lower-amplitude N400s than unrelated words (Kutas, 1993; Federmeier and Kutas, 1999; Ito et al., 2016).

2.2 Cognitive Plausibility of RNN-LMs in N400 modeling

To disentangle the effect of linguistic input from other factors affecting N400 amplitude, a valid model of such linguistic input is needed. Recurrent Neural Network Language Models (RNN-LMs) are, in many ways, perfect models of the 'long-standing and recent linguistic [...] inputs' (Kutas and Federmeier, 2011, p. 641) thought to impact N400 amplitude. Long-standing linguistic inputs in humans are made up of previous language experience, which is analogous to a model's training data; and recent linguistic input is the linguistic context that impacts how humans understand the current utterance, which is analogous to the word sequence preceding the word to be predicted in the model's test data.

Beyond being largely developed as models of human language comprehension (Elman, 1990), recurrent neural network language models (RNN-LMs) have certain properties that make them reasonable models of human cognition. Keller (2010) identifies five features of the human language processing system that he argues are vital for a language model to be cognitively plausible. Three of these are exemplified by unidirectional RNN-LMs—like humans, they can make *predictions* about upcoming words, have a distance-based *memory cost*, and process language word-by-word in order in an *incremental* fashion (unlike bidirectional RNN-LMs and most transformer networks). The two remaining features, *efficiency and robustness* and *broad coverage* are determined more by the model's specific architecture and training than general architecture.

2.3 Surprisal and N400 amplitude

As discussed in Section 2.1, the neurolinguistic evidence suggests that the N400 is a measure of lexical processing difficulty. Recent work, both theoretical and experimental (e.g. Hale, 2001; Levy, 2008; Boston et al., 2008; Demberg and Keller, 2008; Smith and Levy, 2008; Roark et al., 2009; Brouwer et al., 2010; Mitchell et al., 2010; Monsalve et al., 2012; Fossum and Levy, 2012; Frank and Thompson, 2012; Smith and Levy, 2013; Frank, 2014; Willems et al., 2016; Delaney-Busch et al., 2017), has argued that surprisal, the negative logarithm of the probability of a word w_i given its preceding context $w_1 ... w_{i-1}$, as shown in Equation (1), is a good predictor of lexical processing difficulty.

$$S(w_i) = -\log P(w_i | w_1 ... w_{i-1}) \quad (1)$$

Several researchers (Frank et al., 2015; Delaney-Busch et al., 2017; Aurnhammer and Frank, 2019) have directly demonstrated that surprisal is corre-

lated with N400 amplitude. In their study, Delaney-Busch et al. (2017) use a Bayesian approach to calculate the surprisal associated with a target word given a related or unrelated prime (using word association norms and word frequency), and find that this is correlated with N400 amplitude. Frank et al. (2015) and Aurnhammer and Frank (2019) used a number of language models (including RNN-LMs) to calculate the surprisal of words in a natural language text, and compared this to the N400 elicited by these words in human participants, finding a statistically significant correlation.

Frank et al. (2015) and Aurnhammer and Frank (2019) also find that surprisal is a better predictor of N400 amplitude than a number of RNN-LM-derived metrics based on the full probability distributions predicted by the model such as entropy. We suggest that this may be explained by the aforementioned finding that while the N400 amplitude for a word decreases when its semantic content has been pre-activated, it does not increase when a specific prediction is violated. In other words, N400 amplitude is a kind of positive prediction error—a measure of how not-predicted the target word was. This is what surprisal is by definition—it only takes into account how much the actual target word was predicted and is not affected by the rest of the probability distribution. The other metrics, on the other hand, also take into account the rest of the predicted probability distribution, which does not appear to be reflected in N400 amplitude. Thus, there is a theoretical reason for using surprisal to predict N400 amplitude based on previous neurolinguistics research.

2.4 Predicting N400 effects

An alternative approach, that taken by Ettinger et al. (2016), is to use a language-model-derived metric as an analogue of the N400 and investigate whether experimental manipulations in the stimuli that result in statistically significant differences in N400 amplitude also result in statistically significant differences in the chosen metric. This approach allows researchers to investigate whether the reason for the correlation between the metric and N400 amplitude is in fact the experimental manipulation or some other factor.

This is the general approach that we take in this study; however, rather than focusing on the cosine similarity between the word embedding of target word and the combined embeddings of the previous words in the sentence (Ettinger et al., 2016), we model N400 amplitude as surprisal (following Frank et al., 2015; Delaney-Busch et al., 2017; Aurnhammer and Frank, 2019). Additionally, whereas Ettinger et al.'s (2016) proof-of-concept paper is based on 40 sample sentences from a single study investigating one phenomenon, we use stimuli from eleven experiments (with over 100 sentences each) covering a wide range of phenomena.

2.5 Other Models of N400 amplitude

While a number of other researchers have used neural networks to model specific N400 findings this way (Laszlo and Plaut, 2012; Laszlo and Armstrong, 2014; Rabovsky and McRae, 2014; Cheyette and Plaut, 2017; Brouwer et al., 2017; Rabovsky et al., 2018; Venhuizen et al., 2018; Fitz and Chang, 2019), these studies differ in that these models all have semantic representations as part of their input or are trained to learn to output some form of semantic representation. Thus, these models are also limited to the experiments for which they were trained.

For the same reason, these models can also not be used on their own to disentangle the effects of linguistic input from the semantic knowledge provided to them—this can only be done by comparison to models without this. While two of the studies compare their models to simple recurrent networks (SRNs) trained on the same data (Rabovsky et al., 2018; Fitz and Chang, 2019), these SRNs are not representations of the extent of what is possible with lingusitic input alone—these models are simple (for example, they do not use long short-term memory), and much of the power of RNNs comes from large training datasets (see, e.g., the discussion in Chelba et al., 2013).

Finally, it should be noted that while all of the studies discussed in this section aim to model real N400 effects, only two (Laszlo and Armstrong, 2014; Rabovsky and McRae, 2014) use stimuli from real N400 experiments; in the remaining studies, stimuli are chosen to represent manipulations that studies have found to influence N400 amplitude. Given that the N400 is still not fully understood, it is important to verify that the experimental manipulations investigated actually do elicit the expected N400 effect. For this reason, we only use experimental stimuli provided for published N400 experiments, and compare the effect on surprisal directly to the reported effects on N400 amplitude.

3 Approach, Motivations, and Hypotheses

The aim of this study is to investigate the boundary conditions of using surprisal to model N400 amplitude. While there is evidence that surprisal and N400 amplitude are correlated overall (Frank et al., 2015; Aurnhammer and Frank, 2019), it is unclear what variance in N400 amplitude is actually being explained by surprisal. While it is tempting to assume that surprisal is correlated with the N400 because the same factors that lead to reduced N400 amplitudes lead to reduced surprisal, this has thus far not been shown empirically.

This is the question that we investigate in this paper: which experimental manipulations that elicit a difference in N400 amplitude elicit the same difference in surprisal, and which do not?

We do this by running the (English language) stimuli from previously published N400 studies through two neural networks that have been used extensively to model human language processing (e.g., in Wilcox et al., 2018; Futrell et al., 2019; Wilcox et al., 2019; An et al., 2019; Costa and Chaves, 2020). The two models used are the the best English LSTM from Gulordava et al. (2018) and BIG LSTM+CNN INPUTS from Jozefowicz et al. (2016), henceforth (following Futrell et al., 2019) GRNN and JRNN, respectively. These models are both LSTM-RNN-LMs, but differ most notably in size and training data: The JRNN has two hidden layers (8192 and 1024 units), a 793471-word vocabulary, and was trained on 1 billion tokens (Chelba et al., 2013); while the GRNN has two hidden layers (both 650 units), a 50000-word vocabulary, and was trained on 90 million tokens.

In addition to answering questions about the nature of the neurocognitive systems underlying the N400, the results of this study also serve as a baseline for future research—they represent the best that current cognitively plausible neural network language models can do at predicting N400 amplitude using surprisal. Thus, future research that argues for additional sources of information or neurocognitive processes being involved in the N400 on the basis of modeling success should demonstrate that the inclusion of such components in the model improves upon the results presented here.

This aim of establishing a useful baseline is another reason for our choice of models—both are provided pre-trained by the authors, allowing for our results to be replicated and expanded upon. We also only use sets of stimuli that have been made available in papers or their supplementary materials. The stimuli from these papers (Urbach and Kutas, 2010; Kutas, 1993; Ito et al., 2016; Osterhout and Mobley, 1995; Ainsworth-Darnell et al., 1998; Kim and Osterhout, 2005), which cover a range of experimental manipulations that are discussed in Section 4, are included in text format in our supplementary materials[1].

4 Experiments

Figure 1 is a visualization of the findings of the original N400 studies and the results of the simulations. Given the differences in measurements, there is no scale—the heights of the bars indicate which conditions elicited higher or lower N400 amplitudes or surprisals relative to the others in the same experiment or simulation. All and only the significant differences between conditions for significant predictors of the N400 or surprisal are shown, not including significant interactions with recording locations on the scalp (which are beyond the scope of the present study). Black bars represent successful modeling of the differences in N400 amplitude, red bars represent unsuccessful or partially unsuccessful modeling, and purple bars indicate that the results are more complex than can be represented in this way. Only stimuli sets with over 100 stimulus sentences were run through the models (GRNN and JRNN); and while the models were not able to predict the surprisal of all target words (due to limited vocabularies or being unable to process certain characters in sentences), both models successfully calculated the surprisals of over 100 target words in each study. Stimuli, target word surprisals, and the code used to run the models are all included in our supplementary materials.

Where possible, the significant predictors of the surprisal of the GRNN and JRNN models were selected via backwards model selection using likelihood ratio tests of linear-mixed effects models (R Core Team, 2018; Bates et al., 2015) with and without the predictor under investigation as a main effect. When this was not possible, the significance of predictors were evaluated using a Type III ANOVA with Satterthwaite's method for estimating degrees of freedom (Kuznetsova et al., 2017). Significant differences between experimental conditions (i.e. between the levels of a predictor) were

[1] https://github.com/jmichaelov/does-surprisal-explain-n400

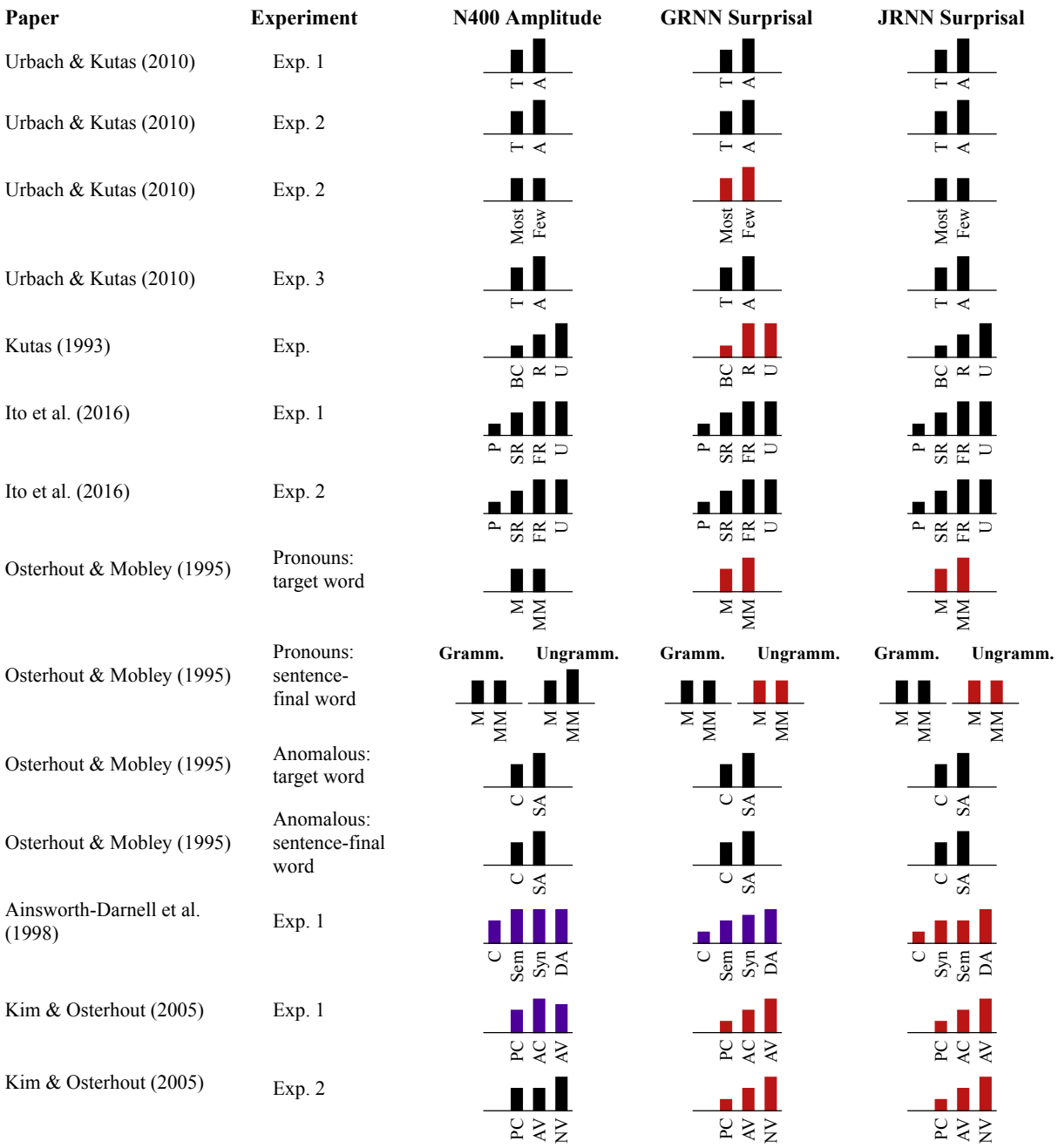

Figure 1: The significant differences between all conditions of significant predictors of N400 amplitude in the original studies and the surprisal of the GRNN and JRNN models. **Black** bars indicate successful modeling of the differences in N400 amplitude, **red** bars indicate unsuccessful or partially unsuccessful modeling, and **purple** bars indicate that the results are more complex than shown.

calculated via t-test based on the selected linear-mixed effects model, using Satterthwaite's method to estimate degrees of freedom (Kuznetsova et al., 2017). In this paper, significant predictors and significant differences between conditions are considered those where $p < 0.05$ in the relevant statistical test. All code for the statistical analyses is included in our supplementary materials.

The remainder of this section discusses the experiments (and the original N400 studies on which they are based) in more detail.

4.1 Urbach and Kutas (2010): Experiment 1

Experiment 1 of Urbach and Kutas (2010) investigates the N400's sensitivity to the typicality of a patient of a described event. There were two kinds of sentences in this experiment exemplified by the following stimulus pair: *prosecutors accuse* ***defendants*** (TYPICAL; T in Figure 1) / ***sheriffs***

(ATYPICAL; A) *of committing a crime*. As expected, the N400 elicited by TYPICAL object nouns is significantly lower in amplitude than that elicited by ATYPICAL object nouns.

Typicality was also a significant predictor of the surprisal of both the GRNN and JRNN models (GRNN: $p < 0.001$; JRNN: $p < 0.001$), with TYPICAL object nouns eliciting a lower surprisal than ATYPICAL ones (GRNN: $p < 0.001$; JRNN: $p < 0.001$).

4.2 Urbach and Kutas (2010): Experiment 2

Expanding on Experiment 1, Urbach and Kutas (2010) ask whether the results are affected by whether the sentences begin with the word *most* or *few* (or synonymous expressions), e.g. ***most*** *prosecutors accuse defendants*. The main effect of typicality remained. In addition, while the main effect of quantifier type was not significant overall (nor was there an interaction with typicality without an interacting electrode location variable), Urbach and Kutas (2010) found that FEW-type quantifiers reduced the N400 amplitude of ATYPICAL patients and reduced the extent to which N400 amplitude was lowered for TYPICAL patients, with this latter effect being found to be statistically significant via t-test.

Typicality predicted the surprisals of both RNNs in the same direction as in Experiemnt 1 ($p < 0.001$ for all statistical tests). The surprisal of the GRNN was also significantly predicted by quantifier type ($p < 0.001$), with FEW-type quantifiers eliciting significantly higher surprisals ($p < 0.001$). As this pattern is limited only to the GRNN (and the analogous main effect does not appear in Experiment 3 for either model), this finding is not considered further. The t-test comparing the N400 of TYPICAL objects under the FEW and MOST quantifiers does not replicate with surprisal—there is no significant difference (GRNN: $p = 0.107$; JRNN: $p = 0.249$).

4.3 Urbach and Kutas (2010): Experiment 3

Experiment 3 of Urbach and Kutas (2010) is a variant of Experiment 2. Instead of MOST or FEW sentence beginnings, the words *often* or *rarely* appear after the subject (agent) noun, e.g. *prosecutors* ***often*** *accuse defendants of committing a crime*. The aim of this was to investigate whether proximity of the quantifier to the target noun had an effect. Urbach and Kutas (2010) again found the same result—only typicality was a significant predictor of N400 amplitude overall; and a t-test found that the N400 reduction for TYPICAL nouns was attenuated by the word *rarely*.

GRNN and JRNN surprisals were only significantly predicted by typicality, with typical nouns eliciting a lower surprisal than atypical nouns ($p < 0.001$ for all tests). The t-test comparing the N400 of TYPICAL objects under the FEW and MOST quantifiers does not replicate with surprisal—there is no significant difference (GRNN: $p = 0.367$; JRNN: $p = 0.283$).

4.4 Kutas (1993)

Kutas (1993) examines the effect of relatedness to the BEST COMPLETION (the highest-cloze completion). An example of a BEST COMPLETION (BC) and RELATED completion can be demonstrated by the following stimulus pair: *The pizza was too hot to* ***chew*** (RELATED; R) / ***eat*** (BC). An example of a BC and UNRELATED pair is the following sentence: *The paint turned out to be the wrong* ***consistency*** (UNRELATED; U)/ ***color*** (BC). BC nouns were found to elicit the lowest N400 amplitude, followed by RELATED nouns, followed by UNRELATED nouns.

Experimental condition is a significant predictor of both GRNN and JRNN surprisal. However, while the surprisals in the GRNN are different between the BC and other nouns ($p < 0.001$ for both RELATED and UNRELATED), there is no significant difference between RELATED and UNRELATED ($p = 0.820$). On the other hand, the surprisals of the JRNN are lowest for BC nouns, followed by RELATED nouns, followed by UNRELATED nouns ($p < 0.001$ for all pairwise comparisons).

4.5 Ito et al. (2016): Experiments 1 and 2

Ito et al. (2016) further investigate the relatedness effect by investigating whether a word that is related in form to the most PREDICTABLE word (i.e. the best completion) has a similar effect on N400 amplitude as being semantically related. The conditions can be illustrated with the following example sentence: *The student is going to the library to borrow a* ***book*** (PREDICTABLE; P)/ ***hook*** (FORM-RELATED; FR)/ ***page*** (SEMANTICALLY RELATED; SR)/ ***sofa*** (UNRELATED; U) *tomorrow*. In both Experiments 1 and 2, where the difference was in the amount of time that the stimuli were presented, Ito et al. (2016) found that experimental condition was a significant predictor, and specifically that PREDICTABLE words elicited the lowest N400 amplitude, followed by SEMANTICALLY RE-

LATED words, followed by the FORM-RELATED and UNRELATED completions, which did not differ in N400 amplitude.

We found the same pattern in the surprisal of both models ($p < 0.001$ for condition as a predictor; $p < 0.001$ for all significant pairwise comparisons; FR vs. U with GRNN surprisal: $p = 0.080$; FR vs. U with JRNN surprisal: $p = 0.399$).

4.6 Osterhout and Mobley: Experiment 2

4.6.1 Pronoun Matching

Osterhout and Mobley (1995) investigate the effect on the amplitude of the N400 elicited by words in sentences where pronouns either do or do not match a preceding noun, as illustrated in the following example: *The aunt heard that **she*** (MATCH; M) / ***he*** (MISMATCH; MM) *had won the lottery*. The MISMATCH sentences can be interpreted as grammatical sentences where the pronoun refers to a different person than that denoted by the sentence subject; or ungrammatical sentences, where the pronoun refers back to the sentence subject with the wrong gender. Osterhout and Mobley (1995) ask whether there is a difference in N400 amplitude between the two conditions, and whether this is affected by which interpretation is taken by participants.

Target Words First, Osterhout and Mobley (1995) look at the N400 measured at the pronoun itself, finding no significant effect of condition.

For both RNN-LMs, however, experimental condition is a significant predictor of surprisal, with matched pronouns eliciting a significantly lower surprisal ($p < 0.001$ for all tests).

Sentence-Final Words The N400 was also measured at the last word in the sentence. Under this condition, it was found that there was a reduced N400 for matching compared to mismatching pronouns, but only for participants who interpreted mismatching sentences to be ungrammatical.

In both models, condition was not found to be a significant predictor of surprisal (GRNN:$p = 0.775$; JRNN: $p = 0.112$). However, whether this is a successful replication of the responses of the participants who found the sentence to be grammatical ('Gramm.' in Figure 1) or a failure to replicate the results of those who found the sentence ungrammatical ('Ungramm.' in Figure 1) is unclear without further research, and thus this result is not discussed further in this paper.

4.6.2 Semantic Anomaly

In parallel to the pronoun stimuli, Osterhout and Mobley (1995) also compared N400 responses to sentences under the following experimental conditions: *The boat sailed down the river and **sank*** (CONTROL; C) / ***coughed*** (SEMANTICALLY ANOMALOUS; SA) *during the storm*.

Target Words N400 amplitude was significantly lower in response to the experimentally manipulated CONTROL words compared to SEMANTICALLY ANOMALOUS words. This effect was replicated in the surprisals of both models ($p < 0.001$ for all tests).

Sentence-Final Words The N400 and surprisals to sentence-final words followed the same pattern as target words ($p < 0.001$ for all tests).

4.7 Ainsworth-Darnell et al. (1998)

Ainsworth-Darnell et al. (1998) investigate the difference in N400 amplitude in response to syntactic and semantic anomaly, operationalized in the following way: *The chef entrusted the recipe **to relatives** before he left Italy* (CONTROL; C) / *The chef entrusted the recipe **to carrots** before he left Italy* (SEMANTIC ANOMALY; SEM) / *The chef entrusted the recipe **relatives** before he left Italy* (SYNTACTIC ANOMALY; SYN) / *The chef entrusted the recipe **carrots** before he left Italy* (DOUBLE ANOMALY; DA). While previous research argued that the N400 does not respond to SYNTACTIC ANOMALY, they found that the CONTROL nouns elicited lower N400 amplitudes than nouns in other conditions, but they did not find a significant difference between the SYNTACTIC ANOMALY and SEMANTIC ANOMALY conditions or between the SEMANTIC ANOMALY and DOUBLE ANOMALY conditions. Ainsworth-Darnell et al. (1998) do not report a test comparing the SYNTACTIC ANOMALY and DOUBLE ANOMALY conditions, but it should be noted that SYNTACTIC ANOMALY has a lower amplitude (based on the graphs) than SEMANTIC ANOMALY, so an unreported significant difference between these should not be ruled out.

Experimental condition is a significant predictor of both GRNN and JRNN surprisal ($p < 0.001$). For both models, the surprisal is lower for words in the CONTROL condition compared to other conditions ($p < 0.001$ for all pairwise comparisons), and there is no significant difference between word in the SYNTACTIC ANOMALY and SEMAN-

TIC ANOMALY conditions (GRNN: $p = 0.274$; JRNN: $p = 0.056$). The surprisals of the two models differ in that while DOUBLE ANOMALY words differ from SEMANTIC ANOMALY words in both models (GRNN: $p < 0.001$; JRNN: $p < 0.001$), they do not differ from the SYNTACTIC ANOMALY in GRNN surprisal but they do in JRNN surprisal (GRNN: $p = 0.059$; JRNN: $p < 0.001$). Based on these findings and inspection of the graphs in Ainsworth-Darnell et al. (1998), it appears that syntactic anomaly of this kind has a larger relative effect on surprisal than N400 amplitude.

4.8 Kim and Osterhout (2005): Experiment 1

Experiment 1 Kim and Osterhout (2005) investigate whether words that violate the event-structure of the described event are still facilitated if they are related to the event being described. The stimuli were of the following form: *The murder had been **witnessed** in the dark* (PASSIVE CONTROL; PC) / *The bystanders had been **witnessing** the crime* (ACTIVE CONTROL; AC) / *The murder had been **witnessing** by the three bystanders* (ATTRACTION VIOLATION; AV). General analysis found that condition only marginally predicted N400 amplitude, but pairwise comparison found one significant difference bwetween conditions: PC completions elicited lower-amplitude N400s than AC completions.

In both models, condition was a significant predictor of surprisal, and PCs elicited the lowest surprisals, followed ACs, followed by AVs ($p < 0.001$ for all tests).

4.9 Kim and Osterhout (2005): Experiment 2

Experiment 2 added the NO-ATTRACTION VIOLATION (NV) condition to the study, which is exemplified by the following sentence: *The unpleasant cough syrup was **witnessing** in the dark*. These were compared to results of the PC and AV conditions in Experiment 1. There was a significant main effect of condition, with PCs and AVs eliciting significantly lower-amplitude N400s than NVs.

Condition was a significant predictor the surprisals of both RNNs, with PCs eliciting a lower surprisal than AVs, followed by NVs with the highest surprisals ($p < 0.001$ for all tests).

5 General Discussion

We compared human N400 responses with surprisal in two RNN-LMs presented with the same stimuli, in the interest of determining the extent to which exposure to linguistic input alone can account for this particular component of human language processing. The results confirmed previous findings that surprisal is generally a good predictor of N400 amplitude, while also clearly demonstrating limitations of the models at capturing the human behavior.

5.1 Successful Predictions

The models effectively predicted certain kinds of contrast that the N400 is sensitive to.

Cloze The surprisals of both models for the Kutas (1993) and Ito et al. (2016) studies show that the surprisal of a language model is sensitive to cloze probability in the same direction as N400 amplitude—higher-cloze words elicit lower N400 amplitudes than lower-cloze words, and the same is true of surprisal.

Relatedness The results of the Kutas (1993) and Ito et al. (2016) experiments also show that surprisal matches N400 amplitude in that words that are related to the highest-cloze completion in terms of semantics, but not form, elicit a lower surprisal than semantically unrelated words, even controlling for these words' cloze.

Semantic typicality The surprisals of both models to the stimuli from Urbach and Kutas's (2010) three experiments demonstrate that the surprisal of a language model patterns in the same way as N400 amplitude in that more typical words (in a given context) elicit a lower surprisal than atypical words in the same context.

Semantic anomaly While the results are framed in the opposite direction in the original studies, the results from the Anomaly stimuli from Osterhout and Mobley (1995) and Experiment 1 of Ainsworth-Darnell et al. (1998) show that, all else being equal, completions that are not semantically anomalous (labeled 'controls' in these experiments) elicit a lower surprisal from language models than semantically anomalous completions, which is the result reported for N400 amplitude in the original studies.

Event structure violations The results for Experiment 2 of Kim and Osterhout (2005) show that both surprisal and N400 amplitude are reduced when a word is in line with event-structure norms, compared to a word that is not and is semantically unrelated to the preceding context.

5.2 Limitations and further directions

At the same time, there are areas where the predictive capabilities of the models are limited.

Quantifiers While the surprisal of the models matched the significant differences in Experiments 2 and 3 of Urbach and Kutas (2010) based on typicality overall, it did not replicate the finding that N400 amplitude was less reduced for TYPICAL nouns when they appeared with FEW or RARELY quantifiers. Thus, it may be the case that some more explicit (or at least more specific) representation of quantification is involved in the neurocognitive processes underlying the N400 than can be modeled by surprisal alone.

Event structure violations Overall, the surprisal of both models is more sensitive to morphosyntactic or event structure violations than N400 amplitude is (for a discussion on the extent to which these can be considered separate in the context of ERPs, see Kuperberg, 2016). For the stimuli from both Kim and Osterhout (2005) experiments, despite the ATTRACTION VIOLATION stimuli eliciting both a significantly reduced N400 amplitude and surprisal compared to the NO-ATTRACTION VIOLATION stimuli, surprisal remained significantly higher for ATTRACTION VIOLATION stimuli than either of the control stimuli, which is not the case with N400 amplitude. Thus, by contrast with the case of quantifiers discussed above (Urbach and Kutas, 2010), which seems to require a more detailed semantic representation, shallower or broader semantic representation might be needed to capture responses to the kinds of stimuli presented in Kim and Osterhout (2005). If the goal is to improve the extent to which models capture human behavior, then there might be ways to accomplish this. Frank and Willems (2017), for example, use cosine distance between the sum of the vectors of all the preceding words in the sentence and the target word to predict the BOLD response (using fMRI) in N400 areas. Given the collateral facilitation of words semantically related to the highest-cloze completions of sentences, it is not unreasonable to assume that a similar process of spreading activation may occur for the preceding as well as the predicted upcoming word in the sentence. One way to implement this could be to weight the RNN model's predictions of the next word by each word's similarity to a general sentence-vector such as that used by Frank and Willems (2017) before the probabilities are transformed into surprisal[2].

Morphosyntactic Anomaly While there has been some discussion about the extent to which event structure violation and morphosyntactic anomalies can be considered separate in the context of ERPs (see, e.g. Kuperberg, 2016), there are clear cases where the surprisal of the language models appear to be more sensitive to morphosyntactic anomaly than N400 amplitude is. This can be seen in humans in the results of Experiment 1 of Ainsworth-Darnell et al. (1998), where words that exhibit either semantic or syntactic anomalies elicit equally reduced surprisal. By contrast, the models predict grammatical continuations to a sentence over ungrammatical ones. This leads to lower surprisals for semantically anomalous words that are syntactically acceptable than those that are both syntactically and semantically anomalous. This difference between humans and the models supports the idea that there needs to be some way to weight predictions by semantic relatedness to the preceding context.

6 Conclusions

Previous work has found that surprisal is a good predictor of N400 amplitude overall. Comparisons of surprisal in RNN-LMs to human N400 responses to the same input sentences showed for the first time that suprisal manages to account for a wide range of phenomena found in human N400 experiments. But at the same time, there are linguistic phenomena where it overpredicts, and others where it underpredicts a significant difference in the human N400 response. From the perspective of human language processing, this suggests that the activation of semantic and lexical features indexed by the N400 cannot be entirely captured by exposure to linguistic input alone. Specifically, quantification, aspects of event structure, and morphosyntactic anomalies seem to require some other learning architecture than the bottom-up statistical learning represented by standard recurrent neural networks. From the perspective of model-building, in order to improve a language-model based cognitive model of the N400, we need to allow for the addition of more shallow semantic processing (independent of syntax and event structure) such as an implementation of spreading activation.

[2]See Kuperberg's (2016) discussion on bag-of-word approaches to the N400.

References

Kim Ainsworth-Darnell, Harvey G Shulman, and Julie E Boland. 1998. Dissociating brain responses to syntactic and semantic anomalies: Evidence from event-related potentials. *Journal of Memory and Language*, 38(1):112–130.

Aixiu An, Peng Qian, Ethan Wilcox, and Roger Levy. 2019. Representation of Constituents in Neural Language Models: Coordination Phrase as a Case Study. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2888–2899, Hong Kong, China. Association for Computational Linguistics.

Christoph Aurnhammer and Stefan L Frank. 2019. Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, 134:107198.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Paul A Bloom and Ira Fischler. 1980. Completion norms for 329 sentence contexts. *Memory & cognition*, 8(6):631–642.

Marisa Ferrara Boston, John Hale, Reinhold Kliegl, Umesh Patil, and Shravan Vasishth. 2008. Parsing costs as predictors of reading difficulty: An evaluation using the potsdam sentence corpus. *Journal of Eye Movement Research.-ISSN*, 2(1):1–12.

Harm Brouwer, Matthew W. Crocker, Noortje J. Venhuizen, and John C. J. Hoeks. 2017. A Neurocomputational Model of the N400 and the P600 in Language Processing. *Cognitive science*, 41:1318–1352.

Harm Brouwer, Hartmut Fitz, and John CJ Hoeks. 2010. Modeling the noun phrase versus sentence coordination ambiguity in dutch: evidence from surprisal theory. In *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics*, pages 72–80. Association for Computational Linguistics.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Samuel J Cheyette and David C Plaut. 2017. Modeling the N400 ERP component as transient semantic overactivation within a neural network model of word comprehension. *Cognition*, 162:153–166.

Jillian K Da Costa and Rui P Chaves. 2020. Assessing the ability of Transformer-based Neural Models to represent structurally unbounded dependencies. In *Proceedings of the Society for Computation in Linguistics (SCiL)*, volume 3, page 10.

Nathaniel Delaney-Busch, Emily Morgan, Ellen F Lau, and Gina R Kuperberg. 2017. Comprehenders rationally adapt semantic predictions to the statistics of the local environment: a bayesian model of trial-by-trial N400 amplitudes. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

Katherine A DeLong and Marta Kutas. 2020. Comprehending surprising sentences: sensitivity of post-N400 positivities to contextual congruity and semantic relatedness. *Language, Cognition and Neuroscience*, pages 1–20.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Allyson Ettinger, Naomi Feldman, Philip Resnik, and Colin Phillips. 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.

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.

Hartmut Fitz and Franklin Chang. 2019. Language ERPs reflect learning through prediction error propagation. *Cognitive Psychology*, 111:15 – 52.

Victoria Fossum and Roger Levy. 2012. Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of the 3rd workshop on cognitive modeling and computational linguistics*, pages 61–69. Association for Computational Linguistics.

Stefan Frank and Robin Thompson. 2012. Early effects of word surprisal on pupil size during reading. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34.

Stefan L. Frank. 2014. Modelling reading times in bilingual sentence comprehension. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

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.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. 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 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. *arXiv:1803.11138 [cs]*. ArXiv: 1803.11138.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.

Aine Ito, Martin Corley, Martin J Pickering, Andrea E Martin, and Mante S Nieuwland. 2016. Predicting form and meaning: Evidence from brain potentials. *Journal of Memory and Language*, 86:157–171.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the Limits of Language Modeling. *arXiv:1602.02410 [cs]*.

Frank Keller. 2010. Cognitively plausible models of human language processing. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 60–67, Uppsala, Sweden. Association for Computational Linguistics.

Albert Kim and Lee Osterhout. 2005. The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language*, 52(2):205–225.

Gina R. Kuperberg. 2016. Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience*, 31(5):602–616.

Gina R Kuperberg, Trevor Brothers, and Edward W Wlotko. 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.

Marta Kutas. 1993. In the company of other words: Electrophysiological evidence for single-word and sentence context effects. *Language and cognitive processes*, 8(4):533–572.

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.

Marta Kutas and Steven A. Hillyard. 1980. Reading Senseless Sentences: Brain Potentials Reflect Semantic Incongruity. *Science*, 207(4427):203–205.

Marta Kutas and Steven A. Hillyard. 1984. Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947):161–163.

Alexandra Kuznetsova, Per B Brockhoff, and Rune Haubo Bojesen Christensen. 2017. lmertest package: tests in linear mixed effects models. *Journal of Statistical Software*, 82(13).

Sarah Laszlo and Blair C Armstrong. 2014. PSPs and ERPs: Applying the dynamics of post-synaptic potentials to individual units in simulation of temporally extended Event-Related Potential reading data. *Brain and language*, 132:22–27.

Sarah Laszlo and David C Plaut. 2012. A neurally plausible parallel distributed processing model of event-related potential word reading data. *Brain and language*, 120(3):271–281.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Steven G. Luke and Kiel Christianson. 2016. Limits on lexical prediction during reading. *Cognitive Psychology*, 88:22–60.

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. Association for Computational Linguistics.

Irene Fernandez Monsalve, Stefan L Frank, and Gabriella Vigliocco. 2012. Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–408. Association for Computational Linguistics.

Lee Osterhout and Linda A Mobley. 1995. Event-related brain potentials elicited by failure to agree. *Journal of Memory and Language*, 34(6):739–773.

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.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Milena Rabovsky, Steven S. Hansen, and James L. McClelland. 2018. Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, 2(9):693–705.

Milena Rabovsky and Ken McRae. 2014. Simulating the N400 ERP component as semantic network error: Insights from a feature-based connectionist attractor model of word meaning. *Cognition*, 132(1):68–89.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 324–333. Association for Computational Linguistics.

Nathaniel J Smith and Roger Levy. 2008. Optimal processing times in reading: a formal model and empirical investigation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 30.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433.

Thomas P. Urbach and Marta Kutas. 2010. Quantifiers more or less quantify on-line: ERP evidence for partial incremental interpretation. *Journal of Memory and Language*, 63(2):158–179.

Cyma Van Petten and Barbara J. Luka. 2012. Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, 83(2):176–190.

Noortje J. Venhuizen, Matthew W. Crocker, and Harm Brouwer. 2018. Expectation-based Comprehension: Modeling the Interaction of World Knowledge and Linguistic Experience. *Discourse Processes*, 0(0):1–27.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN Language Models Learn about Filler–Gap Dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

Ethan Wilcox, Peng Qian, Richard Futrell, Miguel Ballesteros, and Roger Levy. 2019. Structural Supervision Improves Learning of Non-Local Grammatical Dependencies. 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 3302–3312, Minneapolis, Minnesota. Association for Computational Linguistics.

Roel M Willems, Stefan L Frank, Annabel D Nijhof, Peter Hagoort, and Antal Van den Bosch. 2016. Prediction during natural language comprehension. *Cerebral Cortex*, 26(6):2506–2516.