The Role of Context Gating in Predictive Sentence Processing

Yasemin Gokcen (ygokcen@ucmerced.edu)
Department of Cognitive and Information Science, 5200 North Lake Rd.
Merced, CA 95343

David Noelle (dnoelle@ucmerced.edu)
Department of Cognitive and Information Science, 5200 North Lake Rd.
Merced, CA 95343

Rachel Ryskin (rryskin@ucmerced.edu)
Department of Cognitive and Information Science, 5200 North Lake Rd.
Merced, CA 95343

Abstract

Prediction is a core computation in language, as humans use preceding context to implicitly make predictions about the upcoming word in a sentence. In order to do this, humans need some memory for context that is selective and adaptive. We take inspiration from the existing prefrontal cortex literature, in which computational models feature a biologically plausible gating mechanism that can actively maintain and rapidly update task-relevant information to improve performance on cognitive flexibility and working memory tasks. Here, we investigate the potential role of such gating mechanisms in maintaining context for prediction during real-time language processing. Using EEG data from a naturalistic story listening task, we first replicate previous findings that words of low predictability based on preceding context (high in surprisal) elicit larger N400 effects than predictable words. To study how gating may play a role in next-word prediction, we use a performance difference metric between language models with and without gating, which we show is sensitive to word-by-word working memory demand. We find that this gating metric is correlated with EEG amplitude in several later time windows after word onset, providing suggestions concerning the time course of context gating.

Keywords: predictive processing; computational modeling; EEG; neurolinguistics

Introduction

Humans predict upcoming language input based on the preceding sentence context and their prior knowledge of the language and the world (Altmann & Mirković, 2009; Dell & Chang, 2014; Federmeier, 2007; Kuperberg & Jaeger, 2016; Levy, 2008; Ryskin & Nieuwland, 2023, inter alia). For example, the N400 event-related potential (ERP) component — a negative-going deflection of the EEG that peaks around 400 ms after the onset of a meaningful stimulus in centro-posterior channels — is more negative in response to a word that is less predictable from the preceding sentence context compared to one that is more predictable (e.g., S. L. Frank, Otten, Galli, & Vigliocco, 2015; Kutas & Federmeier, 2000). However, the memory representation of the preceding context is likely to be imperfect (e.g., Futrell, Gibson, & Levy, 2020; Lewis & Vasishth, 2005; McElree, Foraker, & Dyer, 2003). Recent evidence suggests that these memory representations are resource-rational — given the constraints of human working memory, they are optimized to minimize downstream surprise (Hahn, Futrell, Levy, & Gibson, 2022). At a computational level of description, each word in the received context has some probability of being retained, and this is combined with prior knowledge of the statistics of the language to generate a distribution over candidate contexts, in turn resulting in a distribution over next-word predictions. Yet, how this resource-rational allocation of memory resources might be approximated algorithmically or neurally is an open question.

Neural network models of working memory (WM) and cognitive control have relied on an adaptive gating mechanism to capture behavior in tasks requiring the selective maintenance and updating of contextual information (Cohen & Servan-Schreiber, 1992; Kriete, Noelle, Cohen, & O'Reilly, 2013; R. O'Reilly & Frank, 2006). For example, in the AX continuous performance task, subjects are presented with a continuous serial stream of stimuli, and they are to respond to one particular stimulus (X), but only if it is immediately preceded by another particular stimulus (A). Thus, a WM for the previous stimulus is needed, and models have shown how networks can learn when to rapidly update the contents of this WM to record a previous stimulus and when to actively maintain the WM contents without disturbing them (Braver & Cohen, 2000). Further, structures that could support this gating function have been found in prefrontal cortex (PFC) (M. J. Frank, Loughry, & O'Reilly, 2001; R. C. O'Reilly, Munakata, Frank, Hazy, & Contributors, 2024), lending biological plausibility to this mechanism.

In this work, we explore the possibility that an adaptive gating mechanism supports the resource-rational allocation of memory resources to context during linguistic prediction. We first replicate the finding that surprisal (an information-theoretic formalization of predictability as $-\log P(word|context)$; Levy, 2008) predicts the amplitude of the N400 ERP component during comprehension, extending it to auditory comprehension using the Natural Stories corpus (Futrell et al., 2021) and extending past work on that corpus by examining processing at high temporal resolution.[1] Here, we are using surprisal as a metric for evaluating predictive processing during naturalistic listening. We then compare word-by-word surprisal estimates using the Natural Stories corpus from two neural network language models with distinct architectures: a recurrent neural network (RNN;

[1] To our knowledge, the Natural Stories corpus has not been utilized in EEG experiments previously.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

Elman, 1990) and a long short term memory network (LSTM; Hochreiter & Schmidhuber, 1997). Crucially, LSTM models incorporate a gating mechanism analogous to those used in neuro-cognitive models of working memory, whereas RNN models do not. We derive a residual measure that captures the extent to which the LSTM models outperform the RNN models. We show that this residual measure is associated with the word-by-word engagement of WM (as operationalized by multiple theories), suggesting that a language model's ability to gate words in the context is particularly helpful in the same situations where WM appears to be taxed in humans. Finally, we explore the neural timecourse of these residual gating effects, finding that they appear to have most predictive power in a time window following the N400.

Methods

Participants

We collected data from healthy young adults (n = 32, mean 21.65 years old with SD of 3.66, 19 female). One subject was excluded due to a lack of EEG triggers for proper analysis. Subjects provided written informed consent and demographic data including language experience, vision and hearing abilities, previous EEG participation, and education. We recruited both monolinguals and bilinguals with the criteria that English must be one of their native languages if they are bilingual (i.e., English learned before the age of 5 years).

Procedure & Materials

Participants listened to stories from the Natural Stories corpus (Futrell et al., 2021). The corpus contains a total of 10 stories, including both nonfiction and fairytales. Each story is between 4-6 minutes in duration. 16 subjects listened to all 10 stories. 15 subjects only listened to 5 randomly selected stories to avoid fatigue. For the subjects who listened to all 10 stories, the order of the stories was either ascending or descending, given the canonical order in the corpus. A fixation cross appeared on the screen for the entire duration of each story. After each story, subjects read 6 yes/no comprehension questions and used a button box to record their response. Participants were also given the option to take small breaks in between stories.

EEG Acquisition

Sixty-four channel EEG data were recorded using Brain Products actiCHamp Plus and BrainVision software systems. Reference and ground electrodes were placed at Cz and AFz, respectively. Additional electrodes were placed on the left and right canthi, below the right eyes, and on the chin. The initial sampling rate was 5000 Hz and target electrode impedance level was set to 25 kOhms.

EEG Data Preprocessing

EEG data were preprocessed using EEGLAB (Delorme & Makeig, 2004). We referenced the data to the average of TP9 and TP10, and applied an IIR Butterworth bandpass with a cutoff frequency of 0.1-30 Hz and 2nd order filter to remove low frequency drifts and high frequency noise. To prepare for independent component analysis (ICA), a copy of the data was made and a filter between 1 and 30 Hz with a 12th order filter was then applied to that copy. No channels were interpolated. We applied the ICA weights from the copied data to the original dataset pre-filtering to then identify and remove one to two components; mostly for eye blinks but occasionally for horizontal eye movements and other noise. Components representing such artifact-related activity were visually identified and removed from the data.

Epochs for each story were extracted time-locked to the onset of the story. Each story was then segmented into word-based epochs time-locked to the onset of each word. The word-based epochs were baselined relative to -300 ms to 0 ms prior to word onset. The timecourse data were downsampled by averaging in 10 ms bins.

Language Models

The RNN and LSTM models from van Schijndel and Linzen (2021) were trained using the RNN modules from PyTorch (Paszke et al., 2019). The models were trained on the Wikitext-2 corpus (Merity, Xiong, Bradbury, & Socher, 2016) to predict the next word, where each model was presented one word at a time. The networks were used to produce surprisal estimates for each word presentation in the Natural Stories corpus. The normalized activation of the output unit corresponding to the correct prediction was taken as the model's estimate of the posterior probability of the target word given previously presented context, and this probability was used to calculate surprisal. We also examined the RNN under two different activation functions: Tanh and ReLU. RNNs have been found to succeed using various activation functions, while the standard LSTM cell incorporates a specified activation function.

Results

For each subject, we examined the accuracy for the post-story comprehension questions as an attention check during story listening. We only omitted stories in which accuracy was 50% or lower. Of 31 subjects, only two subjects had stories with a score below 50%. One subject had one of five stories omitted, while the other had two of ten stories omitted. All other subjects had scored more than 50% on all stories they had listened to, thus no stories were omitted.

Replication of Surprisal Effects on the N400

For each word in the corpus (not including stop words), epochs between -300 to 1000 ms relative to word onset were extracted for the 16 centro-parietal electrodes (Cz, C1, C2, C3, C4, CPz, CP1, CP2, CP3, CP4, Pz, P1, P2, P3, P4). Surprisal values from OpenAI's GPT-Davinci model[2] were taken from the Natural Stories corpus repository. For visualization, words were categorized into three surprisal categories of equal size: high, medium, and low surprisal. For clarity, and

[2]The window size for this model is 2048 tokens.

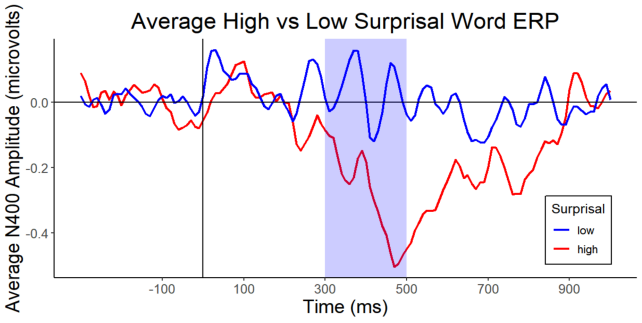

Figure 1: The average timecourse for high versus low surprisal words across all subjects. The x-axis is time relative to the onset of the word where 0ms is the word onset and the y-axis is the average amplitude of the centro-parietal electrodes in μvolts. The blue rectangle highlights the time window used for analysis of the N400 effect, 300-500 ms.

following S. L. Frank et al. (2015), we excluded the medium surprisal group from visualizations. Note that analyses were performed using the continuous surprisal values on all of the data. Figure 1 shows the ERPs for the high and low surprisal words averaged across all subjects and electrodes.

Within the typical N400 time window, the average amplitude for the centro-parietal electrodes in the high surprisal words is more negative relative to the low surprisal words. This pattern extends past the 300-500 ms window, which is often observed in studies using auditory stimuli (Praamstra & Stegeman, 1993; Hagoort & Brown, 2000).

To examine the relationship between amplitude in the N400 time window and surprisal, as well as semantic similarity of the current word to the preceding word, we fit multiple linear mixed-effect models using the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015) in `R`. The word-by-word N400 value was calculated by averaging the amplitude across the 300-500 ms time window. In the full model, the primary predictors of interest included word frequency, semantic similarity between the current word and the previous word, and surprisal values. Semantic similarity was estimated using the cosine similarity between the word's GloVE embedding and the preceding word's embedding, which has been shown in previous work to explain independent variance in N400 amplitudes (Michaelov, Bardolph, Van Petten, Bergen, & Coulson, 2024). Its inclusion helps account for semantic facilitation effects that are not captured by surprisal alone, improving the specificity of inferences about predictive processing. We also included the value of each predictor for the previous word as well as the word's duration, position within the sentence, and position within the story as control predictors. In addition, we included random intercepts for stories and participants and random by-participant slopes for similarity and surprisal. The `lme4` syntax for this model was as follows:[3]

$$\begin{aligned} &meanN400 \sim Frequency + PreviousFrequency \\ &+CosineSimilarity + PreviousCosineSimilarity \\ &+GPTSurprisal + PreviousSurprisal + WordDuration \\ &+SentencePosition + StoryPosition + (1|Story) \\ &+(1 + CosineSimilarity + GPTSurprisal||Subject) \end{aligned}$$

The results are summarized in Table 1. Replicating prior findings, the cosine similarity between the word of interest and the previous word, as well as surprisal, significantly predict N400 amplitude.

To verify that these effects were robust to issues of multicollinearity, we fit simplified models in which one predictor of interest was "ablated" from the fixed effects. The results of model comparison of each ablated model to the full model are summarized in Table 2. Removing cosine similarity and surprisal both significantly reduced model fit.

Table 1: Fixed effects from full model predicting average N400 amplitude

	β	SE	t-value	p-value
Intercept	-0.42	0.30	-1.40	0.16
Frequency	0.05	0.04	1.12	0.26
Prev. Freq.	0.05	0.04	1.28	0.20
Cosine Sim.	0.51	0.20	2.46	0.01*
Prev. Cos. Sim.	-0.22	0.20	-1.10	0.28
GPT Surprisal	-0.04	0.02	-2.62	0.01*
Prev. Surprisal	-.007	0.01	0.51	0.61
Duration	-0.10	0.003	-0.32	0.75
Sentence Position	0.0003	0.003	-0.09	0.93
Story Position	0.0003	0.0003	0.7	0.48

[3]We originally used a more complex model with maximal random effects structure. Due to convergence and singularity issues, we are using a simplified model including random slopes for the two critical predictors of interest only. We calculated VIF scores for all fixed effects using the `car` package in `R` (Fox & Weisberg, 2019) and found all VIF scores to be less than 5, indicating no issues regarding multicollinearity.

Table 2: Results of Model Comparisons between full and ablated models of mean N400 amplitude

Model	Ablated	ΔAIC	ΔBIC	$X^2(1)$	$\Pr(>X^2)$
Full	–	–	–	–	–
Model 2	Freq.	1	10	1.25	0.26
Model 3	Prev. Freq.	1	10	1.61	0.20
Model 4	Cos. Sim.	-3	6	5.53	0.02*
Model 5	Prev. Cos.	1	10	1.30	0.26
Model 6	GPT Surp.	-5	4	6.93	0.008**
Model 7	Prev. Surp.	2	11	0.21	0.64

A Metric of Gating Utility

To investigate the role of context gating in real-time linguistic prediction, we developed a novel metric aimed at capturing when gating is likely to improve prediction of the next word.

Multiple iterations of the RNN and LSTM models, across multiple batch sizes (40 and 128), numbers of hidden units (200, 400, and 650), and learning rates, were trained in order to obtain the version of each model with the best performance on next-word prediction based on cross-entropy loss at test. The LSTMs used in the analysis had the following parameters: batch size of 128 samples, 650 hidden units, one hidden layer, and learning rate of 20. The RNNs used in the analysis had the following parameters: batch size of 128 samples, 650 hidden units, one hidden layer, and learning rate of 0.2. Neither model had a fixed window size, as both models learned temporal relationships within the data itself, unlike feedforward neural networks.[4] To ensure any results were not due to random weights, we trained and tested five iterations of each model type, each with a different random seed, resulting in five LSTM models, five RNN Tanh models, and five ReLU models.

Surprisal values for each word presentation in the Natural Stories corpus were obtained from each of the 15 models. Surprisal values from each LSTM model were regressed onto surprisal values from each RNN model. Figure 2 shows one such regression. As expected, the LSTM and RNN models made similar predictions. Surprisal values from one model robustly predict surprisal values from the other model (βs > 0.88, ps < 0.001, average $R^2 = 0.83$). Surprisal values from the LSTM models were either similar to or somewhat lower than those from RNN models, consistent with the typically observed better next-word prediction performance of LSTM models (e.g., van Schijndel & Linzen, 2018; Aurnhammer & Frank, 2019; Wilcox, Gauthier, Hu, Qian, & Levy, 2020). RNN models (Tanh vs. ReLU) were highly correlated with one another (average r = .949). We focused primarily on the comparison of the LSTMs with both RNNs in order to highlight performance differences with gating, regardless of activation function.

The residual for each word appearance — the difference between the observed LSTM surprisal value and the fitted value predicted from the RNN surprisal value by the linear model — was extracted to serve as a metric of gating utility. Negative residuals represent specific word appearances in which the LSTM is producing lower surprisal values than is predicted from the RNN surprisal values by the linear model. We propose that these are cases where the presence of gating (in the LSTM but not the RNN) facilitated next-word prediction. Positive residuals represent specific word appearances in which the LSTM is producing higher surprisal values than predicted from the RNN surprisal values by the linear model. We propose that these are cases where the presence of gating (in the LSTM but not the RNN) hampered (or minimally, didn't facilitate) next-word prediction.

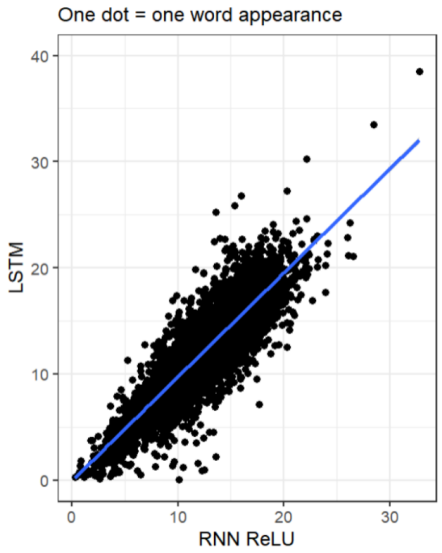

Figure 2: Correlation between LSTM and RNN ReLU surprisal values. Each point indicates a particular appearance of a word. The x-axis is the RNN ReLU surprisal value for that word appearance and the y-axis is the LSTM surprisal value. The gating utility residual value for each word is the vertical distance to the regression line in blue. Dots below the line are cases in which the LSTM is producing lower surprisal values than is predicted by the regression model based on the RNN surprisal values (negative residuals) — where "gating utility" is high.

This process was repeated for every combination of LSTM-to-RNN Tanh and LSTM-to-RNN ReLU, resulting in 50 regressions and, therefore, 50 sets of gating utility residuals.

To test the external validity of this residual-based gating utility metric, we compare these residuals to word-by-word WM demand measures. If residuals index words for which the presence of gating, an important mechanism implicated in WM, is important for accurate prediction, we would expect a negative correlation between these residuals and a measure which captures the need for/engagement of WM (i.e., as WM demand increases, gating utility increases).

WM demand during language processing has been the topic of long-standing theoretical debate and development in psycholinguistics. For the purposes of validating the gating utility metric, we use all of the word-by-word measures of WM demand available in Shain, Blank, Fedorenko, Gibson, and Schuler (2022). These measures of WM demand spans three separate frameworks: Dependency Locality Theory (DLT) (Gibson, 2000), Adaptive Character of Thought—Rational (ACT-R) (Anderson et al., 2004), and Left-Corner Parsing (LCP) (Johnson-Laird, 1983; Lewis & Vasishth, 2005). DLT measures linguistic complexity based

[4]The two model architectures were equated by setting the stated parameters (except learning rate) to be equal. There may be other techniques to equate the models that we did not explore.

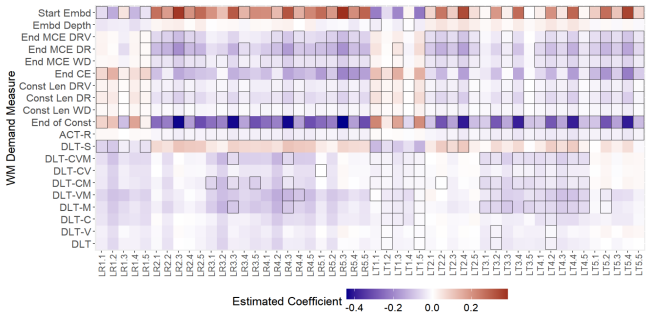

Figure 3: Heatmap of the 500 regressions between the residual-based gating utility metric and 20 measures of WM demand. Each cell of the raster plot represents one regression between a set of word-by-word residuals and WM demand measures. The x-axis of the raster plot shows one tick for each set of residuals defined by the LSTM and RNN which were used to compute residuals. (L =LSTM, R=ReLU, T=Tanh. The first number is the LSTM iteration and the second is the RNN iteration, such that LR3.2 is LSTM Iteration 3 vs. RNN ReLU Iteration 2.) The y-axis of the raster plot denotes the WM demand measure. The color of the cell indicates the size and directionality of the β from each regression (purple is more negative, red is more positive). Cells that are outlined in black represent a statistically significant relationship based on Bonferroni corrected α levels.

on the distance between words forming a dependency, with greater distances increasing WM demand. Variants of DLT account for factors like whether the word is a verb, a modifier, or coordinates with another word. ACT-R models WM retrieval costs rather than storage, assuming words are retrieved only when needed. It incorporates memory decay, meaning word representations fade over time but can be retrieved, and similar consecutive words result in lower WM demand. LCP, like DLT, measures the distance between words, but focuses on the number of constituents or word embeddings in a sentence. WM demand increases as more embeddings and constituents are processed simultaneously, with variants considering word distance, verb status, and discourse referents.

Twenty linear models — one per WM demand measure — were fit to each set of 50 residuals[5], resulting in 500 regressions. Figure 3 summarizes the βs and p-values of these regressions. Bonferroni correction was applied to the p-values, based on the WM theory. The adjusted alpha level used for the DLT variables was $\alpha = 0.0055$ and $\alpha = 0.01$ for the embedding/constituent variables.

Despite variability across the various model combinations, most of the regressions yielded negative βs, meaning that, generally, across theoretical frameworks for capturing WM demand, as WM demand increased, residual values became more negative. This indicates that the presence of gating may

[5]This includes 25 LSTM-to-Tanh residuals and 25 LSTM-to-ReLU residuals

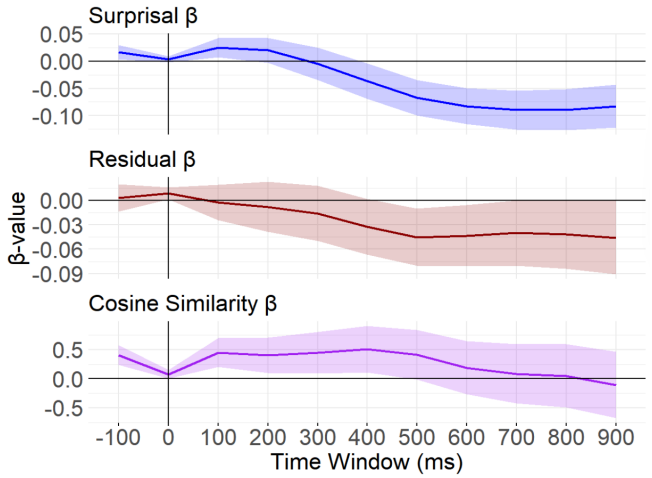

Figure 4: The β coefficient of surprisal, the LSTM-RNN Residual, and Cosine Similarity Value across the word epoch. The midpoint of each time window is shown on the x-axis and the β coefficient value is show in the y-axis. The ribbon indicates the 95% confidence interval.

support prediction in contexts where WM demands are high. In the cases (most reliably in the Start Embedding measure) where the relationship was significant and positive, as WM demand increased residuals became more positive, indicating that the presence of gating did not facilitate prediction. This does not necessarily mean that gating actively hindered prediction, however.

Sliding-Window Gating Effect Analysis

To investigate when the gating mechanism may be involved in real-time next-word prediction, we conducted a sliding-window analysis using the previously discussed gating utility metric.

Using the same set of electrodes as for the analysis of the N400 effect, we compute the average amplitude per word in a series of 200 ms time windows starting with -200 to 0 ms relative to word onset and ending with 800 to 1000 ms, in steps of 100 ms (i.e., -200 to 0 ms, -100 to 100 ms, etc.). In each of these time windows, we fit a linear mixed-effects model that is nearly identical to the base model used in the analysis of the N400 effect but adding the gating utility (residual) measure as a fixed effect and random slope. The inclusion of many relevant linguistic predictors helps isolate the unique contribution of gating utility, reducing the likelihood that the effect is driven by confounding factors.

Figure 4 summarizes the results of this analysis for the three predictors of most theoretical interest: surprisal, gating residual, and cosine similarity. The effect of surprisal is significantly negative in all time windows starting with the 300-500 ms time window, consistent with the prior analysis of the N400 time window. (It is also positive in two consecutive early time windows -100-100 ms and 0-200 ms.) The effect of the gating utility residual is significantly negative in

the two consecutive time windows 400-600 ms and 500-700 ms, as well as 800-1000 ms. It is also significantly positive in the -200-0 ms. The effect of cosine similarity is significantly positive in all consecutive time windows until 300-500 ms (inclusive). Other effects not in this graph include word audio duration, which is only significantly positive during the 800-1000 ms time window. No other effects were significant in any time windows.

Discussion

We report results from a new dataset of EEG during story listening with the Natural Stories corpus. Using this corpus, we replicate previous findings that words with higher surprisal elicit more negative N400 ERP components (e.g., S. L. Frank et al., 2015; Michaelov & Bergen, 2020; Szewczyk & Federmeier, 2022; Brouwer, Delogu, Venhuizen, & Crocker, 2021). The semantic similarity of the current word to the preceding word also explains the independent variance in the amplitude of N400, consistent with previous work (e.g., S. L. Frank & Willems, 2017; Franklin, Dien, Neely, Huber, & Waterson, 2007; Levari & Snedeker, 2024).

We also proposed and tested a metric of gating utility by comparing performance on next-word prediction between LSTM models, which incorporate a gating mechanism, and RNN models, which do not. This metric appears to correlate with increases in WM demand. Situations in which WM demand is high are indeed those in which we would expect that the presence of gating would be most facilitatory for next-word prediction. This extends previous work by leveraging the residual difference as a word-by-word metric of WM demand, and linking it directly to neural activity during naturalistic comprehension.

As a first step toward understanding the role of gating in real-time prediction, we used a sliding-window analysis to explore the timecourse of the effect of our gating utility metric on ERP amplitude. We observed a significant negative effect starting after the N400 time window and continuing for several hundreds of milliseconds. As the utility of gating for the current word increases (the residual becomes more negative), the ERP amplitude becomes more positive. We can speculate that using gating to maintain or update the memory representation for the context leads to more accurate prediction and this positivity reflects the processing consequences of a "poorly gated" context memory representation (i.e., one that was lossy and lead the listener to predict something different than what they received). It is interesting to note that the time window overlaps with the time window commonly used to define the P600 ERP component, which has been tied to the engagement of WM/cognitive control (Ovans et al., 2022) during language processing. Further investigation will be necessary to understand the relationship between gating and the P600. While we take these results as evidence of a physiological signal that scales with the use of WM during prediction, we remain agnostic about the specific WM mechanisms causally related to N400 amplitude. These questions will be explored in future modeling work aimed at formally linking WM computations, such as gating, and ERP dynamics.

References

Altmann, G. T. M., & Mirković, J. (2009). Incrementality and Prediction in Human Sentence Processing. *Cognitive Science*, *33*(4), 583–609. doi: 10.1111/j.1551-6709.2009.01022.x

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–1060. doi: 10.1037/0033-295X.111.4.1036

Aurnhammer, C., & Frank, S. L. (2019). Comparing gated and simple recurrent neural network architectures as models of human sentence processing. In *The 41st annual conference of the cognitive science society.*

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

Braver, T. S., & Cohen, J. D. (2000). On the control of control: The role of dopamine in regulating prefrontal function and working memory. *Attention and Performance*, *18*, 712–737.

Brouwer, H., Delogu, F., Venhuizen, N. J., & Crocker, M. W. (2021). Neurobehavioral Correlates of Surprisal in Language Comprehension: A Neurocomputational Model. *Frontiers in Psychology*, *12*. doi: 10.3389/fpsyg.2021.615538

Cohen, J. D., & Servan-Schreiber, D. (1992). Context, cortex, and dopamine: A connectionist approach to behavior and biology in schizophrenia. *Psychological Review*, *99*(1), 45. doi: 10.1037/0033-295X.99.1.45

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), 20120394–20120394. doi: 10.1098/rstb.2012.0394

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. doi: 10.1016/j.jneumeth.2003.10.009

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211. doi: 10.1207/s15516709cog1402_1

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*(4), 491–505. doi: 10.1111/j.1469-8986.2007.00531.x

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (Third ed.). Thousand Oaks CA: Sage. Retrieved from `https://www.john-fox.ca/Companion/`

Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001). Interactions between frontal cortex and basal ganglia in work-

ing memory: A computational model. *Cognitive, Affective & Behavioral Neuroscience*, *1*(2), 137–160. doi: 10.3758/cabn.1.2.137

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. doi: 10.1016/j.bandl.2014.10.006

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

Franklin, M. S., Dien, J., Neely, J. H., Huber, E., & Waterson, L. D. (2007). Semantic priming modulates the N400, N300, and N400RP. *Clinical Neurophysiology*, *118*(5), 1053–1068. doi: 10.1016/j.clinph.2007.01.012

Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing. *Cognitive Science*, *44*(3), e12814. doi: 10.1111/cogs.12814

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S. T., & Fedorenko, E. (2021). The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, *55*(1), 63–77. doi: 10.1007/s10579-020-09503-7

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, language, brain: Papers from the first mind articulation project symposium* (pp. 94–126). The MIT Press.

Hagoort, P., & Brown, C. M. (2000). ERP effects of listening to speech: Semantic ERP effects. *Neuropsychologia*, *38*(11), 1518–1530. doi: 10.1016/S0028-3932(00)00052-X

Hahn, M., Futrell, R., Levy, R., & Gibson, E. (2022). A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, *119*(43), e2122602119. doi: 10.1073/pnas.2122602119

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press.

Kriete, T., Noelle, D. C., Cohen, J. D., & O'Reilly, R. C. (2013). Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proceedings of the National Academy of Sciences*, *110*(41), 16390–16395. doi: 10.1073/pnas.1303547110

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59. doi: 10.1080/23273798.2015.1102299

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, *4*(12), 463–470. doi: 10.1016/s1364-6613(00)01560-6

Levari, T., & Snedeker, J. (2024). Understanding words in context: A naturalistic EEG study of children's lexical processing. *Journal of Memory and Language*, *137*, 104512. doi: 10.1016/j.jml.2024.104512

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. doi: 10.1016/j.cognition.2007.05.006

Lewis, R. L., & Vasishth, S. (2005). An Activation-Based Model of Sentence Processing as Skilled Memory Retrieval. *Cognitive Science*, *29*(3), 375–419. doi: 10.1207/s15516709cog0000_25

McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of memory and language*, *48*(1), 67–91.

Merity, S., Xiong, C., Bradbury, J., & Socher, R. (2016). *Pointer Sentinel Mixture Models* (No. arXiv:1609.07843). arXiv. doi: 10.48550/arXiv.1609.07843

Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., & Coulson, S. (2024). Strong Prediction: Language Model Surprisal Explains Multiple N400 Effects. *Neurobiology of Language*, *5*(1), 107–135. doi: 10.1162/nol_a_00105

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). doi: 10.18653/v1/2020.conll-1.53

O'Reilly, R., & Frank, M. (2006). Making Working Memory Work: A Computational Model of Learning in the Prefrontal Cortex and Basal Ganglia. *Neural Computation*, *18*(2), 283–328.

O'Reilly, R. C., Munakata, Y., Frank, M. J., Hazy, T. E., & Contributors. (2024). *Computational cognitive neuroscience*. Online Book, 5th Edition, URL: `https://compcogneuro.org`. Retrieved from `https://compcogneuro.org/book`

Ovans, Z., Hsu, N. S., Bell-Souder, D., Gilley, P., Novick, J. M., & Kim, A. E. (2022). Cognitive control states influence real-time sentence processing as reflected in the P600 ERP. *Language, Cognition and Neuroscience*, *0*(0), 1–9. doi: 10.1080/23273798.2022.2026422

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library* (No. arXiv:1912.01703). arXiv. doi: 10.48550/arXiv.1912.01703

Praamstra, P., & Stegeman, D. F. (1993). Phonological effects on the auditory N400 event-related brain potential. *Cognitive Brain Research*, *1*(2), 73–86. doi: 10.1016/0926-6410(93)90013-U

Ryskin, R., & Nieuwland, M. S. (2023). Prediction during language comprehension: What is next? *Trends in Cognitive Sciences*, *27*(11), 1032–1052. doi:

10.1016/j.tics.2023.08.003

Shain, C., Blank, I. A., Fedorenko, E., Gibson, E., & Schuler, W. (2022). Robust Effects of Working Memory Demand during Naturalistic Language Comprehension in Language-Selective Cortex. *Journal of Neuroscience*, *42*(39), 7412–7430. doi: 10.1523/JNEUROSCI.1894-21.2022

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. doi: 10.1016/j.jml.2021.104311

van Schijndel, M., & Linzen, T. (2018). A Neural Model of Adaptation in Reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4704–4710). Brussels, Belgium: Association for Computational Linguistics. doi: 10.18653/v1/D18-1499

van Schijndel, M., & Linzen, T. (2021). Single-Stage Prediction Models Do Not Explain the Magnitude of Syntactic Disambiguation Difficulty. *Cognitive Science*, *45*(6), e12988. doi: 10.1111/cogs.12988

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). *On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior* (No. arXiv:2006.01912). arXiv. doi: 10.48550/arXiv.2006.01912