Neural language model gradients predict event-related brain potentials

Stefan L. Frank
Centre for Language Studies
Radboud University
Nijmegen, the Netherlands
stefan.frank@ru.nl

Abstract

Fitz and Chang (2019) argue that event-related brain potentials during sentence comprehension result from the detection and incorporation of word-prediction error. Specifically, the N400 component would correlate with prediction error while the P600 component would be indicative of error backpropagation in the language system. The current work evaluates this hypothesis on a corpus of EEG data recorded during naturalistic sentence reading. Word-prediction error and backpropagated error were estimated by an LSTM language model that processed the same 205 English sentences as the human participants. At each word, the word's surprisal and the total gradient of recurrent-layer connections were collected for comparison to the sizes of the N400 and P600 components. Consistent with the theory, higher surprisal resulted in stronger N400 while higher gradient resulted in stronger P600, and ERPs on content words were more sensitive to surprisal whereas ERPs on function words were more sensitive to gradient. However, a detailed analysis of the neural signal's time course indicates that the apparent P600 effect could be interpreted as a reversed N400 effect.

1 Introduction

1.1 Event-related brain potentials

When people engage in language comprehension, their brains display particular patterns of electrical activity, a small part of which can be picked up by electrodes on the scalp. This method, known as electroencephalography (EEG), has revealed several typical deflections in measured voltage in response to word perception. These deflections are known as event-related brain potentials (ERPs) and particular ERP components can be identified by their timing and scalp distribution.

Arguably the two most studied components are the N400 and the P600. The first of these components is a negative-going voltage deflection that peaks at around 400 ms after word onset, hence the name N400. The second ERP component goes in the positive direction and peaks at around 600 ms after word onset, hence the name P600.

The N400 is well know to be stronger (i.e., more negative) on words that are syntactically correct but semantically odd (Kutas and Hillyard, 1980), or simply have lower occurrence probability as estimated by human judgements (Kutas and Hillyard, 1984) or language models (Frank et al., 2015). A stronger P600 was originally thought to be indicative of syntactic violations and anomalies (Osterhout and Holcomb, 1992) but has also been found in different types of well-formed sentences, for example in response to a word that completes a long-distance dependency (Kaan et al., 2000) or is used ironically (Regel et al., 2014).

1.2 Models of the N400 and P600

Several computational models have been proposed as explanations of N400 and P600 effects in language comprehension (Brouwer et al., 2017, 2021; Fitz and Chang, 2019; Li and Futrell, 2022, 2023). These models agree that the N400 is stronger in response to a word that was less expected, although they differ in how this prediction error is quantified. As for the P600, all models assume that its size correlates with the extent to which the incoming word results in an update of some representation, but they disagree on the content of this representation.

According to the Retrieval-Integration account by Brouwer et al. (2017, 2021), the P600 corresponds to the update of a representation of the situation described by the sentence or text, which in their model is represented at the output layer of a recurrent neural network. In contrast, the model by Li and Futrell (2022, 2023) assumes that the P600 reflects the update in the reader's (or listener's) beliefs about the word sequence processed so far. Finally, Fitz and Chang's (2019) Error Propagation account claims that processing a word can lead

Proceedings of the Society for Computation in Linguistics (SCiL) 2024, pages 316-323.
Irvine, California, June 27-29, 2024

to an update of language knowledge, that is, to learning about the language's statistics or syntactic patterns. The P600 would reflect the size of this knowledge update, which can be quantified as the backpropagated word-prediction error in a neural network.

Fitz and Chang (2019) tested their theory in the Dual-Path model (Chang et al., 2006), a recurrent neural network (RNN) that differs from most language models in that it splits processing into two paths: a syntactic path that takes care of word ordering and a semantic path that maps propositional meaning onto sentences. During model training, each word's prediction error backpropagates through both paths but converges on a single recurrent layer. Fitz and Chang (2019) take the summed absolute gradients of recurrent-layer connection weights as their predictor of the P600 induced by the word, and show that this accounts for many results from human P600 experiments, such as the stronger P600 response to syntactic violations (compared to grammatically correct alternatives) caused by subject-verb number disagreement or incorrect verb-tense inflections. More recently, Verwijmeren et al. (2023) demonstrated that the Error Propagation account, implemented in a bilingual version of the Dual-Path model (Tsoukala et al., 2021) can explain why subject-verb number disagreement results in an enhanced N400 in beginning second-language learners but an enhanced P600 in more advanced learners.

1.3 The current study

In spite of its successes, evaluation of the Error Propagation account has been hampered by the limitations of training the Dual-Path model, which requires each sentence to be paired with its propositional semantics. In practice, this means that the model can only be trained on artificial, toy versions of real languages. Although this often suffices for investigating specific psycholinguistic phenomena, it makes broad-coverage validation on natural language impossible. Crucially, Fitz and Chang (2019) did also investigate the Dual-Path model's ERP predictions when the semantic path was removed, in effect reducing it to a normal simple recurrent network (Elman, 1990) trained on the same toy language as the full Dual-Path model. Results were similar to those of the full model (at least, as far as the P600 was concerned) suggesting that the semantic path contributed little, if anything, to the P600 prediction.

If semantic knowledge is indeed not required to explain P600 effects, the Error Propagation account can also be evaluated in a way that is more similar to common practice in computational linguistics: train a neural language model on a natural language corpus and then test it on a novel sample of sentences. This is exactly the approach I take here. An RNN is used to estimate word surprisal and the word-induced gradients of recurrent-layer connection weights, at each word of English sentences that were also read by native English speakers while their EEG was recorded. Next, linear regression predicts the human N400 and P600 sizes from the model-derived surprisal and gradient values.

The results of the current study show that, as predicted by the Error Propagation account, higher surprisal correlates with stronger N400 while higher gradient correlates with stronger P600. Unexpectedly, however, higher surprisal and gradient also correspond to *weaker* P600 and N400, respectively. This suggests that the two predictors in fact have the same effect on the EEG signal (albeit in opposite directions) and the apparent separable effects on the N400 and P600 components are an artifact caused by their spatiotemporal overlap. This interpretation is supported by additional regression analyses: Across time and scalp locations, surprisal and gradient show similar effects on the EEG signal. Hence, backpropagated word-prediction error may thus correspond to weaker N400s as opposed to stronger P600s.

2 Methods

2.1 EEG data

Frank et al. (2015) published EEG data recorded on 32 electrodes, from 24 native English speakers reading 205 English sentences that were extracted from novels. The sentences were presented word-by-word[1] at a fixed location to minimize eye movements that interfere with the EEG signal. The duration between consecutive word onsets was at least 627 ms and increased by 20 ms per character, that is, it was word-length dependent.

Time-locked to each word onset, the EEG signals were averaged over different combinations of scalp electrodes and time windows to obtain six ERP components that have been investigated in the psycho- and neurolinguistic literature (see Frank et al., 2015, for details). The baseline level for each

[1] Punctuation marks were attached to the preceding word and contractions were presented as single words.

component was the average over that component's electrodes during the 100 ms leading up to word onset. Here, I investigate only the N400 and P600 components. The N400 is defined as the average voltage from 300 to 500 ms after word onset, over 12 centro-parietal electrodes. The P600 is the average from 500 to 700 ms after word onset, over 18 electrodes that include the N400 electrodes but also more temporally located ones.

2.2 Language model[2]

2.2.1 Model architecture

As mentioned in Section 1.2, Fitz and Chang (2019) tested their theory in an RNN next-word prediction model that has both a semantic and a syntactic pathway (although they found the semantic pathway not to be critical to the P600 predictions). In order to stay as close as possible to that architecture while allowing it to be trained on a natural language corpus, I sacrificed the semantic pathway, leaving a plain, single-layer RNN; more specifically, a Long Short-Term Memory model (LSTM; Hochreiter and Schmidhuber, 1997) with 400-dimensional input embeddings and a 500-unit recurrent (LSTM) layer followed by a 400-unit hidden layer before the softmax output layer.

2.2.2 Model training

Training sentences were extracted from the first 7 slices of the ENCOW16 corpus of English sentences from the web (Schäfer, 2015). First, a vocabulary was created comprising the 20,000 most frequent tokens in the first slice of ENCOW16 plus all tokens from the 205 experimental stimuli sentences.[3] Next, all sentences were selected that contain only vocabulary tokens and are no less than 3 and no more than 50 tokens in length. This resulted in a total of just under 81.6M training sentences with over 1.4B tokens of 21,918 types. All tokens from the experimental stimuli were attested in this training set. The training set was presented to the network for 1 training epoch.

2.2.3 Model testing

At several points during training, the LSTM processed the 205 sentences from the Frank et al. (2015) EEG study and estimated each word's surprisal (Hale, 2001; Levy, 2008), that is, the negative log-probability of the word conditioned on the sentence so far. Surprisal values quantify word-prediction error and are expected to correlate with the size of the N400 component, as was already shown by Frank et al. (2015) on the same EEG data but using surprisal estimates from much smaller language models.

Each word's prediction error is backpropagated through the network (Rumelhart et al., 1986) resulting in a gradient for each connection weight. Following Fitz and Chang (2019), I take the summed absolute values of the gradients in the recurrent layer; an aggregate measure I simply refer to as 'the gradient'. Unlike Fitz and Chang's (2019) simple recurrent network's units, LSTM units have four types of connection (for the memory cell, and the input, output, and forget gates). The gradient measure is computed over all these weights together. Note that the gradients are computed for the 205 experimental sentences but not actually applied during model testing, that is, the connection weights are not updated.

2.3 Data analysis

Following Frank et al. (2015), I exclude from analysis all sentence-initial words, words attached to punctuation, and any data point from part of the EEG signal that was considered an artifact (mostly due to eye blinks). This left a total of 33,476 data points (i.e., combinations of participants and word tokens) for analysis. Statistical models were fit by the MixedModels package (Bates et al., 2023) in Julia (Bezanson et al., 2017).[4]

2.3.1 Standard ERP analysis

Separate sets of linear mixed-effects regression analyses were run with N400 size or P600 size as the dependent variable. Both analyses included surprisal and gradient as predictors, and the following covariates of no interest: the component's baseline, the position of the sentence in the experiment session, the position of the word in the sentence, the log-transformed frequency of the word in the British National Corpus, and the word's length

[2] The language model's PyTorch (Paszke et al., 2019) code, training data, and trained models can be downloaded from `https://osf.io/a6g4f`

[3] Sentences from another psycholinguistic study were also included but these are irrelevant to the current work. The corpus sentence tokenization was adapted to that of the EEG experiment by merging the parts of a contraction (e.g., the two corpus tokens "do␣n't" become the single token "don't"). Punctuation marks remained individual tokens.

[4] Analysis code and EEG data can be downloaded from `https://osf.io/a6g4f`.

(number of characters). All predictors were standardized. The regression models included a by-token random intercept and slope of sentence position, and a by-participant random intercept and slopes of surprisal, gradient, sentence position, word position, log-transformed word frequency, and word length.

I take the t-statistics of surprisal and gradient as measures of the extent to which they are predictive of ERP size. A negative t-value of surprisal is expected in the N400 analysis (higher surprisal leads to a stronger, i.e., more negative-going N400) and a positive t-value of gradient in the P600 analysis (higher gradient leads to a stronger, i.e., more positive-going P600). When $|t| > 2$, this roughly corresponds to an effect that is statistically significant with $p < .05$.

2.3.2 Regression ERP analysis

A follow-up analysis does not take the ERP sizes as dependent variables but follows the 'regression ERP' (rERP) approach of Smith and Kutas (2015). This comes down to fitting a regression model to the set of EEG samples at each time point (relative to word onset) and electrode, and then plotting the coefficients of the predictors of interest as if they are ERP curves. All these regression models have both surprisal and gradient as predictors, with the same covariates and random-effect structure as in the standard ERP analysis discussed above. To reduce computation time, this analysis is only performed for the 7 most central electrodes, using only the fully trained network's surprisal and gradient estimates.

3 Results

3.1 Surprisal and gradient measures

Figure 1 shows how the per-sentence averages of surprisal and gradient, as well as the correlation between them, change over network training. As expected, surprisal decreases with more training, indicating the the network makes increasingly accurate next-word predictions. Put differently: it is learning the statistical patterns of English.

Perhaps more surprisingly, gradient initially remains low, so not much of the prediction error in the output units results in changes in the recurrent connection weights. After approximately 100K training sentences, prediction error is increasingly backpropagated to the LSTM layer until the gradient more or less stabilizes after 10M sentences.

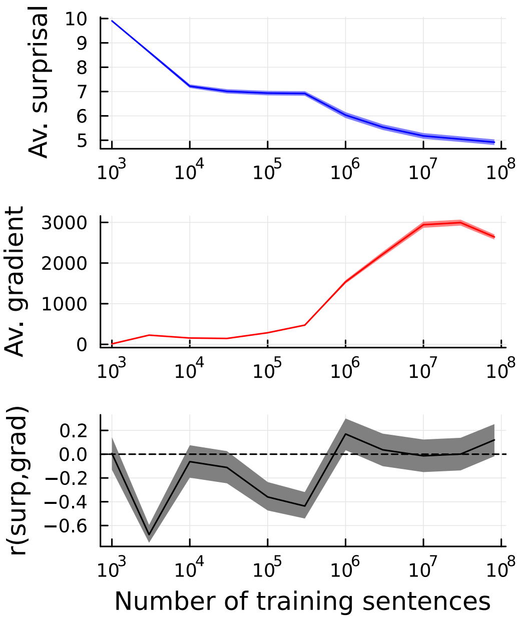

Figure 1: Average surprisal (top), average gradient (center), and their correlation (bottom) as a function of the number of training sentences. Shaded areas indicate 95% confidence intervals. Averages, correlations, and confidence intervals are computed over the 205 per-sentence averages because within a sentence, the word-level values do not constitute independent measurements.

There is a medium-sized, negative correlation between surprisal and gradient until about 300K training sentences, but after the network has been trained on 1M sentences the correlation is no longer statistically significant. The negative correlation early in training may seem hard to reconcile with the fact that output prediction error (quantified by surprisal) is backpropagated and then forms the driving force behind connection weight update (quantified by gradient). I return to this issue in Section 4.2.

3.2 Standard ERP analysis

Figure 2 shows how the fit of surprisal and gradient to ERP size changes as the number of training sentences increases. Clearly, high surprisal leads to a stronger (more negative-going) N400, and this effect of surprisal increases as the model is more thoroughly trained. The effect of gradient on P600 size is weaker, but it is in the positive direction and also increases with more training.

N400 effects are known to be mostly driven by content words (Frank et al., 2015) while the P600 has often been associated with syntactic processing.

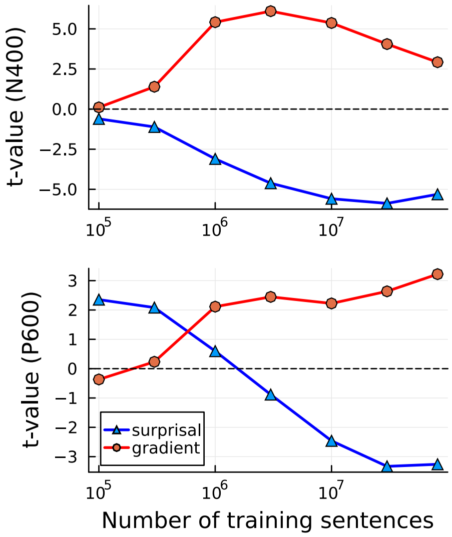

Figure 2: t-statistics for the effects of surprisal (blue triangles) and gradient (red circles) on N400 size (top) and P600 size (bottom), as a function of the number of training sentences.

To investigate if this distinction is apparent in the effects of surprisal and gradient, content and function words were also analyzed separately.[5] As Figure 3 shows, surprisal is more predictive of ERP size on content words than on function words, whereas the same is not the case for gradient.

3.3 Regression ERP analysis

In addition to the expected effects of surprisal and gradient, the standard ERP analysis of Section 3.2 revealed that higher surprisal results in weaker P600 and that higher gradient results in weaker N400 (although the latter effect decreases after about 3M training sentences). This is most likely due to spatio-temporal overlap between the two ERP components (Brouwer and Crocker, 2017), which raises the question whether the apparent P600 effect of gradient truly is a P600 or if it could be a reversed effect on the N400 that only looks like a P600 because the two components are not fully separated in time and electrode location.

The results of the rERP analysis in Figure 4 suggest that this is indeed the case: The positive effect of gradient peaks at around 400 ms instead of 600 ms after word onset.

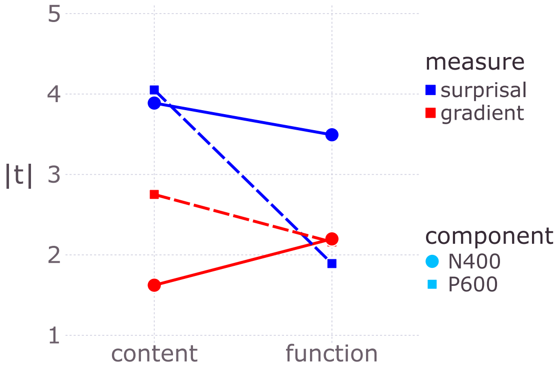

Figure 3: Absolute values of the t-statistics for the effects of surprisal (blue) and gradient (red), after training on the full dataset, analyzed separately for content and function words. Solid lines and round markers denote N400 effects; dashed lines and square markers denote P600 effects.

4 Discussion

4.1 The Error Propagation account

According to the Error Propagation account of language-related ERPs, the N400 during sentence comprehension reflects word-prediction error and the P600 corresponds to the (potential) update in language knowledge caused by word processing. Fitz and Chang (2019) quantify the size of this update in terms of the gradients of recurrent connection weights in an RNN. So far, this hypothesis had only been evaluated by comparing P600-size predictions between pairs of input sentences that constituted 'toy', artificial versions of controlled stimuli from psycholinguistics experiments. The current study, in contrast, is the first to validate the Error Propagation account on EEG data from a naturalistic sentence comprehension experiment, extracting the N400 and P600 predictions from a neural language model trained on a reasonably sized corpus of natural language text.

The results partially support Fitz and Chang's (2019) theory: prediction error (surprisal) is predictive of the N400 and backpropagated error (gradient) corresponds to a positive-going ERP. Also, the finding that only surprisal effects are stronger for content words than for function words (Figure 3) is consistent with the idea that surprisal mainly affects the N400 and gradient the P600. Clearly, surprisal and gradient have separable effects in the expected directions. However, the regression-ERP analysis revealed that what was assumed to be an P600 in fact has a time course that is more like that of an N400 (one that is weaker for higher gradient)

[5] This follows the content/function-word split provided by Frank et al. (2015), where 53.2% of words were designated as content words and 46.8% as function words. Contractions were excluded.

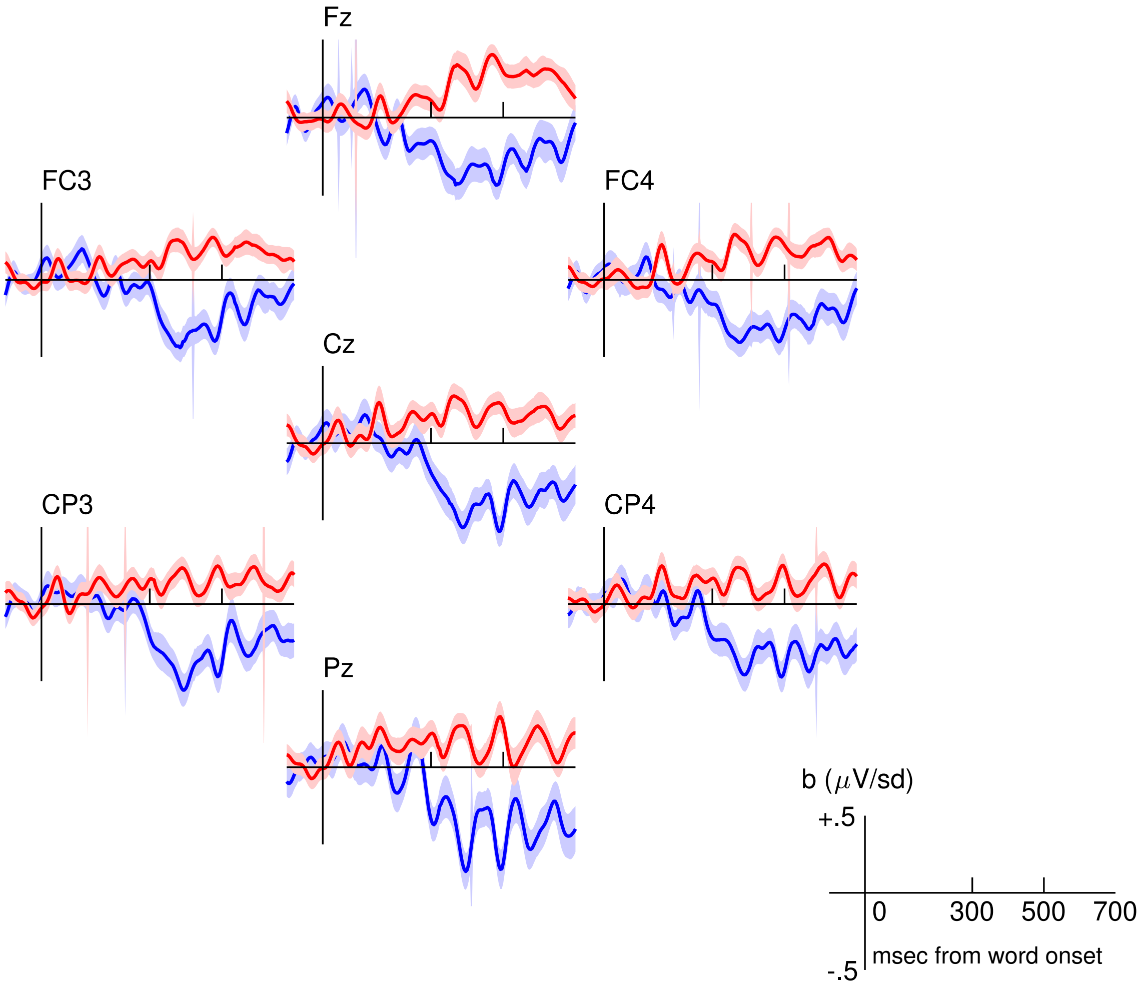

Figure 4: Topographic map of rERP curves. Each plot corresponds to one electrode and the curves show the effects (regression coefficients) of surprisal (blue) or gradient (red) on voltage at the electrode, time-locked to word onset. Shaded areas indicate standard errors.

and may therefore not be a true P600 ERP component. To summarize, the findings are inconclusive: There is an effect of gradient although it may not be exactly the effect predicted by the theory. The question remains whether surprisal and gradient indeed form qualitatively different linking hypothesis between properties of the language model and properties of the EEG signal, or if there are merely two sides of the same coin, with gradient modulating (i.e., weakening) the effect of surprisal on the N400.

4.2 Correlation between surprisal and gradient

Figure 1 revealed an unexpected and fairly large negative correlation between surprisal and gradient during early stages of network training. Although error backpropagation can only occur to the extent that there is prediction error, gradient and surprisal are not simply the same measure because the gradient of a connection's weight also depends on the activation going into that connection. Moreover, there can be confounding variables between surprisal and gradient. Possibly, a confound with word frequency is responsible for the observed negative correlation between surprisal and gradient. Word frequencies will be among the first statistics learned by the network, where they are encoded in the output units' biases. As is visible from Figure 1, at the point in training when surprisal and gradient are negatively correlated, average surprisal has dropped but gradient remains close to 0, indicating that very little prediction error is backpropagated: learning mostly takes place at the output connections and biases. Presumably, the output units representing high-frequency words are the first to have fairly stable biases, so prediction errors on these words are the first to be backpropagated, resulting in non-zero gradients in the LSTM layer. Meanwhile, prediction error on low-frequency words still mostly leads to changes in output biases. As a consequence, gradients in the LSTM layer will be higher on higher-frequency (and, consequently, lower-surprisal) words, that is, the surprisal and gradient measures are negatively correlated.

4.3 Evaluation on experimental versus naturalistic items

P600 effects in sentence comprehension are mostly, if not exclusively, investigated on sentences that result in comprehension difficulty, be it due to (morpho)syntactic violations (Coulson et al., 1998), garden-path structures (Osterhout and Holcomb, 1992), long-distance dependencies (Kaan et al., 2000), or semantic incongruity (Kuperberg et al., 2003). The same is true for all models of ERP effects discussed in Section 1.2. In contrast, the Frank et al. (2015) test sentences were sampled from novels and are therefore not expected (nor manipulated) to evoke any specific difficulty. It is not impossible that for such easy-to-understand sentences, the P600 occurs earlier, coinciding with the N400. Future research may reveal if the Error Propagation account, in combination with a neural language model trained on natural text, predicts more standard P600 effects on the hand-crafted sentences from psycholinguistic experiments.

Note that such an evaluation on realistic data is not possible with the Retrieval-Integration account (Brouwer et al., 2017, 2021) because that account takes the P600 to reflect the update of a representation of the described situation, and therefore requires such a representation – something that is not easily formalized for natural language. In contrast, the Li and Futrell (2022, 2023) model only requires knowledge of syntactic word-order patterns and therefore can be (and, in fact, has been) evaluated using the actual stimuli of psycholinguistic experiments.

4.4 Improving the language model

Another potential avenue for future research is to investigate whether improving the quality of the language model also improves its ERP predictions. The current work stayed as close as possible to that of Fitz and Chang (2019), using a single-layer RNN. Increasing the network's size (e.g., adding layers), changing the architecture (e.g., a Transformer instead of an LSTM), and increasing the amount of training data will certainly result in a more accurate language model. In general, better language models more accurately fit human processing measures, be it from EEG, eye tracking, or fMRI (Merkx and Frank, 2021; Schrimpf et al., 2021). With multiple network layers to extract the gradient measure from, it may also be possible to distinguish between P600s resulting from different aspects of language processing.

5 Conclusion

This study tested Fitz and Chang's (2019) Error Propagation account of event-related brain potentials during sentence comprehension, by extracting N400 and P600 predictions from a neural language model that processed the same sentences as humans in an EEG study. In line with the theory, the model's word-prediction error (surprisal) correlated with N400 size. Backpropagated word-prediction error, which quantifies the potential update of the reader's language knowledge, is measurable in the EEG signal but it remains unclear whether this takes the form of a stronger P600 or a weaker N400.

Acknowledgements

I am grateful to Naomi Shapiro for helpful comments on an earlier version of this paper.

References

Douglas Bates, Phillip Alday, Dave Kleinschmidt, José Bayoán Santiago Calderón, Likan Zhan, Andreas Noack, Milan Bouchet-Valat, Alex Arslan, Tony Kelman, Antoine Baldassari, Benedikt Ehinger, Daniel Karrasch, Elliot Saba, Jacob Quinn, Michael Hatherly, Morten Piibeleht, Patrick Kofod Mogensen, Simon Babayan, Tim Holy, Yakir Luc Gagnon, and Yoni Nazarathy. 2023. Juliastats/mixedmodels.jl: v4.22.3.

Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. 2017. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98.

Harm Brouwer and Matthew W. Crocker. 2017. On the proper treatment of the N400 and P600 in language comprehension. *Frontiers in Psychology*, 8:1327.

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, Francesca Delogu, Noortje J. Venhuizen, and Matthew W. Crocker. 2021. Neurobehavioral correlates of surprisal in language comprehension: A neurocomputational model. *Frontiers in Psychology*, 12:615538.

Franklin Chang, Gary S. Dell, and Kathryn Bock. 2006. Becoming syntactic. *Psychological Review*, 113(2):234.

Seana Coulson, Jonathan W. King, and Marta Kutas. 1998. Expect the unexpected: Event-related brain response to morphosyntactic violations. *Language and Cognitive Processes*, 13(1):21–58.

J. L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14:179–211.

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

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.

John T. Hale. 2001. A probabilistic Early parser as a psycholinguistic model. In *Proceedings of the second conference of the North American chapter of the Association for Computational Linguistics*, volume 2, pages 159–166. Pittsburgh, PA: Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Edith Kaan, Anthony Harris, Edward Gibson, and Phillip Holcomb. 2000. The P600 as an index of syntactic integration difficulty. *Language and Cognitive Processes*, 15(2):159–201.

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

Marta Kutas and Steven A. Hillyard. 1980. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207:203–205.

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

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

Jiaxuan Li and Richard Futrell. 2022. A unified information-theoretic model of EEG signatures of human language processing. In *NeurIPS 2022 Workshop on Information-Theoretic Principles in Cognitive Systems*.

Jiaxuan Li and Richard Futrell. 2023. A decomposition of surprisal tracks the N400 and P600 brain potentials. In *Proceedings of the 45th Annual Meeting of the Cognitive Science Society*.

Danny Merkx and Stefan L. Frank. 2021. Human sentence processing: recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*.

Lee Osterhout and Phillip J. Holcomb. 1992. Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, 31(6):785–806.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pages 8024–8035.

Stefanie Regel, Lars Meyer, and Thoomas C. Gunter. 2014. Distinguishing neurocognitive processes reflected by P600 effects: Evidence from ERPs and neural oscillations. *PLoS ONE*, 9(5):e96840.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Roland Schäfer. 2015. Processing and querying large web corpora with the COW14 architecture. In P. Bański, H. Biber, E. Breiteneder, M. Kupietz, H. Lüngen, and A. Witt, editors, *Proceedings of the 3rd Workshop on the Challenges in the Management of Large Corpora*, pages 28–34. Institut für Deutsche Sprache, Mannheim, Germany.

Martin Schrimpf, Idan A. Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118:e2105646118.

Nathaniel J. Smith and Marta Kutas. 2015. Regression-based estimation of ERP waveforms: I. The rERP framework. *Psychophysiology*, 52:157–168.

Chara Tsoukala, Mirjam Broersma, Antal Van Den Bosch, and Stefan L. Frank. 2021. Simulating code-switching using a neural network model of bilingual sentence production. *Computational Brain & Behavior*, 4:87–100.

Stephan Verwijmeren, Stefan L. Frank, Hartmut Fitz, and Yung Han Khoe. 2023. A neural network simulation of event-related potentials in response to syntactic violations in second-language learning. In *Proceedings of the 21st International Conference on Cognitive Modelling*.