A unified information-theoretic model of EEG signatures of human language processing

Anonymous Author(s)
Affiliation
Address
email

Abstract

We advance an information-theoretic model of human language processing in the brain, in which incoming linguistic input is processed at two levels, in terms of a heuristic interpretation and in terms of error correction. We propose that these two kinds of information processing have distinct electroencephalographic signatures, corresponding to the well-documented N400 and P600 components of language-related event-related potentials (ERPs). Formally, we show that the information content (surprisal) of a word in context can be decomposed into two quantities: (A) **heuristic surprise**, which signals processing difficulty of word given its inferred context, and corresponds with the N400 signal; and (B) **discrepancy signal**, which reflects divergence between the true context and the inferred context, and corresponds to the P600 signal. Both of these quantities can be estimated using modern NLP techniques. We validate our theory by successfully simulating ERP patterns elicited by a variety of linguistic manipulations in previously-reported experimental data from [23]. Our theory is in principle compatible with traditional cognitive theories assuming a 'good-enough' heuristic interpretation stage, but with precise information-theoretic formulation.

1 Introduction

Human language comprehension is linked to (at least) two distinct and robust event-related potential (ERP) components detectable through electroencephalography—the N400 and P600. The N400 is a negative-going waveform that peaks at around 400 ms after the onset of linguistic signal, whereas the P600 is a positivity at around 600 ms. Since their discovery, a great deal of research has attempted to ascertain the functional interpretation of the N400 and P600 signals in order to shed light on the neural mechanisms of human language processing [eg. 16, 9, 11, 13, 25, 26, 14, 15, 26].

Recent psycholinguistic theories have proposed a **heuristic interpretation** stage of language comprehension, where comprehenders form a plausible interpretation based on a subset of information in the input signal [13, 15, 25]. This idea has been formalized in a noisy-channel framework, where comprehenders rationally infer a probabilistic distribution on the intended utterance given the received input while taking into account the fact that the input may contain errors ("noise"). Prior work has established that there is a reduced N400 and a larger P600 when a plausible corrected sentence can be recovered from an original sentence containing a semantic error, and that the P600 amplitude is related to the likelihood function in the noisy-channel framework [8, 23]. However, none of the proposed theories can currently explain the full range of empirical ERP patterns (see [2]), and existing models are not integrated with more general computational neuroscientific models.

We propose an information-theoretic computational-level model of the N400 and P600 ERP components in language processing, formalizing the noisy-channel intuition described above and integrating

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

multiple strands of psycholinguistic research into a quantitative model that explains previously-reported results while making successful novel predictions about linguistic ERPs[1].

2 Model

Our model builds on Surprisal Theory, an empirically successful theory of behavioral signatures of language comprehension such as reading time [10, 17, 3, 24, 27], is in line with recent computational neuroscientific proposals to quantify cognitive effort information-theoretically [21, 28, 7, 12, 4]. Surprisal Theory holds that the magnitude of processing effort for a word x_t given a context of previous words $x_{<t}$ should be proportional to the information content or **surprisal** S_t of the word given its context:

$$S_t = \ln p(x_t \mid x_{<t}). \tag{1}$$

Our model maintains the idea that the total amount of processing effort is given by surprisal, but we partition the surprisal into two parts, corresponding to different forms of information processing and to the two distinct ERP signals.

Surprisal decomposition Consider a comprehender perceiving a sentence at time t, currently observing word x_t in the context of (a memory trace of) previously-observed words $x_{<t}$. We formalize the idea of a 'heuristic interpretation' in the generative model shown in Figure 1. Here the comprehender is trying to infer the value of a variable T representing the speaker's intended structure, for example a complete parse tree. Crucially, the link between the intended structure T and the input words x is not deterministic: speakers may make errors in production, or environmental noise may disrupt the signal, and comprehenders should be able to correct for these factors. We formalize this idea by introducing random variables for **heuristic words** $W_{<t}$ and W_t, corresponding to the values of past words and the current word within the speaker's inteded structure T. The heuristic words give rise to the input words through a **noise model**, a distribution $p_N(x \mid W)$ representing all kinds of errors that might occur during language production and transmission.

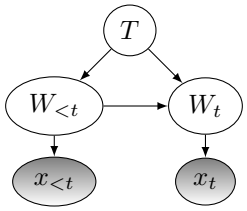


Figure 1: The comprehender's generative model. T is the speaker's intended structure. At time t, the structure T contains words $W_{<t}$ and (the past context) and W_t (current word). The comprehender observes a noisy form of the past context $x_{<t}$ and the current word x_t.

We propose that, with each incoming word, the comprehender is updating her representations of the heuristic words W and structure T. Within the generative model of Figure 1, the surprisal S_t can be partitioned into two parts, corresponding to (A) the new information content of the heuristic words themselves, termed **heuristic surprise**, and (B) the update to beliefs about the heuristic words given the input words, termed **discrepancy signal**:

$$S_t = \underbrace{\langle -\ln p(W_t \mid W_{<t}) \rangle}_{\text{heuristic surprise},=A} + \underbrace{\left\langle \ln \frac{p(W_t \mid W_{<t})}{p(x_t \mid x_{<t})} \right\rangle}_{\text{discrepancy signal},=B}, \tag{2}$$

where $\langle \cdot \rangle$ indicates an average with respect to the probability distribution $p(W_{\leq t} \mid x_{\leq t})$. The heuristic surprise is an upper bound on the information provided by the heuristic words about the structure T. We propose that the N400 magnitude is proportional to the heuristic surprise A and the P600 magnitude is proportional to the discrepancy signal B for distinct positive scalars α and β in:

$$\text{N400} = \alpha A, \text{P600} = \beta B, \tag{3}$$

Noise model The model quantities A and B are both averages with respect to the comprehender's probability distribution on heuristic words given input words, $p(W \mid x)$. This distribution can be written using Bayes' rule as

$$p(W \mid x) \propto p_N(x \mid W) p(W). \tag{4}$$

[1]We make all the codes and data are available at `https://anonymous.4open.science/r/NeurIPS-Workshop-Surprisal-Decomposition-ERP-CF9C/`.

To fully specify the model, therefore, requires us to specify (1) a noise model p_N representing likely errors in production and/or transmission, and (2) a prior probability distribution $p(W)$, which reflects the probability that a speaker would want to produce a sequence of words W.

Implementation We follow Eq. 4 applied independently to individual words. Here $p(W)$ for a single word is calculated using the Masked Language Model RoBERTa [18], and $p(x \mid W)$ for a single word is calculated as

$$p(x \mid w) \propto e^{-\lambda d(x,w)}, \tag{5}$$

where λ is a constant free parameter and $d(w, x)$ is the Levenstein edit distance between input word x and heuristic word w. In order to generate candidate corrections, we replace target word in the input sentence with a special token `<mask>`, and use RoBERTa to generate the probability distribution to fill in the masked token. We selected top 100 predictions as our W. After that, we calculate the posterior probability by multiplying the RoBERTa probability with Eq. 5. We calculate the conditional probability of the current word W_t given the context $W_{<t}$ with the autoregressive transformer GPT-2 [22]. The conditional probability of veridical target x_t given the veridical context $x_{<t}$ is also calculated with GPT-2.

3 Empirical Validation

3.1 Dataset

We validate our model using ERP data from [23], who report experiments designed to test how well noisy-channel error correction can explain linguistic ERP patterns.[2] The experiments have four conditions (see Table 1, one with a semantic violation (*Sem*), one with a syntactic violation (*Syn*), one semantic critical condition (*SemCrit*) with a semantic violation which could be attributed to noise, and a control sentence without any error (*Control*). The ERP effects in all three experimental conditions are all calculated in terms of differences to the ERP signal in the control condition. In the N400 time window, there is a significant N400 effect in *Sem* and *SemCrit* conditions, where the N400 effect in *SemCrit* condition is reduced (see Fig. 3a. In the P600 time window, there is a significant P600 effect in *Synt* and a smaller but significant P600 effect in *SemCrit* condition (see Fig. 3c).

Condition	**Sentence**	**Effect**
Semantic	The storyteller could turn any incident into an amusing *hearse*.	N400
Syntactic	The storyteller could turn any incident into an amusing *anecdotes*.	P600
SemCrit	The storyteller could turn any incident into an amusing *antidote*.	N400, P600
Control	The storyteller could turn any incident into an amusing *anecdote*.	—

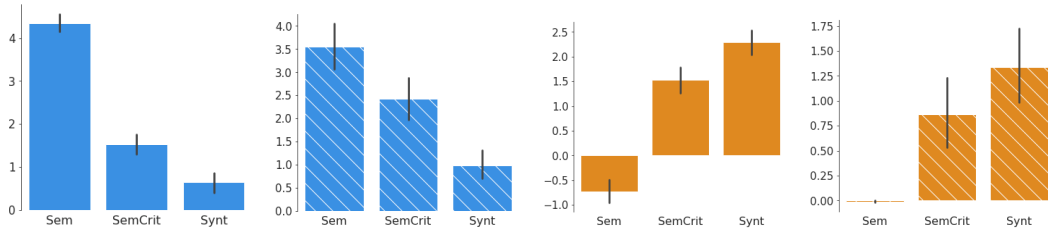
Table 1: List of conditions, sample sentences and ERP effects in dataset.

3.2 Simulation Results

Fig. 2 show the simulated and empirical N400 and P600 effect size across conditions in the dataset, with $\lambda = 400$ (see Appendix A.2 for results of a hyperparameter search on λ). The simulated effect sizes from the RoBERTa-based model implementation are similar to the effect sizes in real human ERP experiments. One potential issue in the use of language models for word probabilities is that, although large language models are sensitive to syntactic violations [5, 6], the surprisal penalty associated with syntactic violations is smaller than for semantic violations (see Appendix A.1). This means that, in order to get corrections in the *Synt* condition, it is necessary to set λ to a very high value, 400. Model results with a smaller value of $\lambda = 300$ are shown in Appendix A.2; these show a better fit to the *SemCrit* condition for the N400.

We statistically confirmed the relationship between the empirical ERP amplitudes (N400 and P600) and our information-theoretic measures (heuristic surprise A and discrepancy signal B) in maximal linear mixed-effects models including by-subject and by-item intercepts and slopes [1]. We find a significant main effect of heuristic surprise on N400 amplitude ($t = -6.57, p < .001$), and a

[2] The data is available at `https://osf.io/vcsfb/` without a license.



(a) N400 (Human) (b) N400 (Simulated) (c) P600 (Human) (d) P600 (Simulated)

Figure 2: N400 and P600 effect size from human experiment and from model prediction ($\lambda = 400$).

significant main effect of discrepancy signal on P600 amplitude ($t = 3.65, p < .001$). In comparison, we find no significant effect of true surprisal on P600 ($t = 0.20, p = 0.84$), suggesting that our proposed decomposition of surprisal provides a better fit of the ERP components than true surprisal [20, 19].

Table 2: The effects of true surprisal S_t, heuristic surprise A, and discrepancy signal B on ERP amplitudes in the N400 and P600 time range in the experiment from [23]. Numbers are β values (t-values). $p < 0.05$*, $p < 0.01$**, $p < 0.001$***. A negative effect in the N400 range indicates the standard N400 effect; a positive effect in the P600 range indicates the standard P600 effect.

N400		P600	
heuristic surprise	true surprisal	discrepancy signal	true surprisal
-0.66 (-6.57***)	-0.59 (-5.76***)	0.75 (3.65***)	0.02 (0.20)

4 Conclusion

We presented a neuro-computational model of N400 and P600 ERP components in language processing. We modeled the ERP components based on a generalized theory of surprisal. We argue that surprisal of word can be decomposed into two parts— a heuristic surprise and a discrepancy signal, which corresponds to N400 and P600 respectively. The two measures have a clear cognitive interpretation. The heuristic surprise signals the processing difficulty of target word given heuristic context, and the discrepancy signal represents the efforts of updating the discourse from inferred to true structure. We approximate the distribution of heuristic interpretation via a noisy-channel process, and implement it with large-scale language models. The theory is validated with experimental results.

Our model provides an information-theoretic quantification of language-related neural signals. By linking ERP components to Surprisal Theory, our model creates a precise formal link between theories of ERPs and other behavioral measures of language processing. The work calls for co-registration of brain and behavioral experiments to better understand the underlying cognitive process.

Our model highlights the role of probabilistic inference in language processing, and provides a computational implementation of it. Our model innovatively generates probabilistic distribution of heuristic inference with a noisy-channel model, without having to fully commit to the cognitive interpretation of the noisy-channel process. The noisy-channel model abstracts away how different linguistic cues are weighted and combined by evaluating the heuristic interpretation based on a balance between prior belief and its divergence with new evidence. Furthermore, we leverage recent computational models from the field of natural language processing to implement our theory, which allows us to take into account the statistical variations in real experimental inputs. While we provide an insightful implementation by quantifying parameters with pre-trained language models and edit distance, we want to acknowledge that this model is on a computational-level with minimal hyper-parameter searching. Further work could be done on the precise nature of the heuristic interpretation generation process.

References

[1] Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278, 2013.

[2] Harm Brouwer, Hartmut Fitz, and John Hoeks. Getting real about semantic illusions: rethinking the functional role of the P600 in language comprehension. *Brain research*, 1446:127–143, 2012.

[3] Stefan L. Frank and Rens Bod. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22(6):829–834, 2011.

[4] Richard Futrell and Michael Hahn. Information theory as a bridge between language function and language form. *Frontiers in Communication*, 7:657725, 2022.

[5] Richard Futrell, Peng Qian, Edward Gibson, Evelina Fedorenko, and Idan Blank. Syntactic dependencies correspond to word pairs with high mutual information. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 3–13, Paris, France, 27–28 August 2019. Association for Computational Linguistics.

[6] Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online, July 2020. Association for Computational Linguistics.

[7] Samuel J Gershman. Origin of perseveration in the trade-off between reward and complexity. *bioRxiv*, 2020.

[8] Edward Gibson, Leon Bergen, and Steven T. Piantadosi. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20):8051–8056, 2013.

[9] Peter Hagoort, Colin Brown, and Jolanda Groothusen. The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and cognitive processes*, 8(4):439–483, 1993.

[10] John Hale. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics*, 2001.

[11] John CJ Hoeks, Laurie A. Stowe, and Gina Doedens. Seeing words in context: the interaction of lexical and sentence level information during reading. *Cognitive brain research*, 19(1):59–73, 2004.

[12] Anthony MV Jakob and Samuel J Gershman. Rate-distortion theory of neural coding and its implications for working memory. *bioRxiv*, 2022.

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

[14] Gina R. Kuperberg. Neural mechanisms of language comprehension: Challenges to syntax. *Brain research*, 1146:23–49, 2007.

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

[16] Marta Kutas and Steven A. Hillyard. Event-related brain potentials to semantically inappropriate and surprisingly large words. *Biological psychology*, 11(2):99–116, 1980.

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

[18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[19] James Michaelov and Benjamin Bergen. How well does surprisal explain N400 amplitude under different experimental conditions? In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 652–663, 2020.

[20] James Michaelov, Seana Coulson, and Benjamin Bergen. So cloze yet so far: N400 amplitude is better predicted by distributional information than human predictability judgements, 2021.

[21] Pedro A. Ortega and Daniel A. Braun. Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A: Mathematical, Physical, and Engineering Sciences*, 469:20120683, 2013.

[22] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[23] Rachel A. Ryskin, Laura Stearns, Marianna Eddy, Leon Bergen, Evelina Fedorenko, and Edward Gibson. The P600 ERP component as an index of rational error correction within a noisy-channel framework of human communication. *BioRxiv*, 2020.

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

[25] Marieke Van Herten, Herman HJ Kolk, and Dorothee J. Chwilla. An ERP study of P600 effects elicited by semantic anomalies. *Cognitive brain research*, 22(2):241–255, 2005.

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

[27] Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912*, 2020.

[28] Alexandre Zénon, Oleg Solopchuk, and Giovanni Pezzulo. An information-theoretic perspective on the costs of cognition. *Neuropsychologia*, 123:5–18, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We propose and test an information-theoretic model of N400 and P600 in language processing.
   (b) Did you describe the limitations of your work? [Yes] See the discussion of future work in the section Conclusion.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This work is concerned with cognitive process of language comprehension and adopts a computational approach. We cannot foresee any potential negative societal impacts
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] see Appendix A.1. Data Availability.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] The section 2 Model and Appendix A2 provide model details.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] The fig 2. includes by-item error bars of empirical results and model predictions.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] The computing can be done on a modern personal computer or google Colab with less than three hours.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [Yes]

   (b) Did you mention the license of the assets? [No] The assets are not licensed.

   (c) Did you include any new assets either in the supplemental material or as a URL? [No]

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] The data are publicly available on `https://osf.io/vcsfb/`

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] The data are either created by researchers or reported on publicly available journals with proper ethics reviews

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

A.1 Comparison between human- and LM-generated word probabilities.

Table 3 shows the cloze probability obtained from human sentence completion task and from GPT-2 generations. All three experimental conditions (*Sem*, *SemCrit* and *Synt*) have close to 0 human cloze probabilities, however, GPT-2 assigns a lower surprisal to the *Synt* condition than the other two semantically anomalous conditions. The discrepancy between human and language model probabilities indicates that language models might have under-estimated the surprisal of syntactically anomalous sentences.

Table 3: Averaged cloze probability and total surprisal of stimuli across conditions.

	Control	*Sem*	*SemCrit*	*Synt*
Human cloze probability	0.40	0	0	0
GPT-2 generated surprisal	5.2	8.7	8.5	7.5

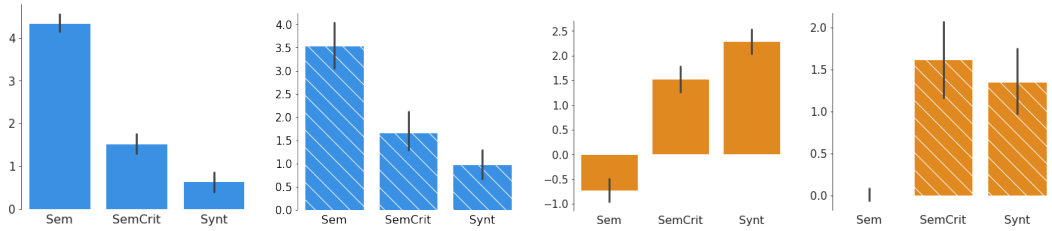
A.2 Hyper-parameter Tuning

We explored the effect of λ with a grid search from 100 to 500, with a step size of 100, and with two marginal conditions ($\lambda = 0$ and $\lambda = 1000$). The simulated N400 and P600 across our selection of λ, together with true surprisal of the stimuli, are summarized in Table 4. When $\lambda = 0$, the heuristic surprise is the surprisal the most predictable words given the previous, regardless of the true target received. As λ increases, it becomes more different to do error correction, with an increased heuristic surprise and a decreased error correction. Importantly, sentences in different conditions have different sensitivity to λ. Sentences in *Synt* and *SemCrit* have an easy fix, and therefore are more likely to be corrected given an increasingly large λ. After visual inspection, we chose $\lambda = 400$.

Table 4: Averaged heuristic surprise and discrepancy signal under different λ values.

	Heuristic Surprise ($\rightarrow$ N400)				Discrepancy Signal ($\rightarrow$ P600)			
	Control	*Sem*	*SemCrit*	*Synt*	*Control*	*Sem*	*SemCrit*	*Synt*
$\lambda = 0$	3.9	4.0	3.9	3.9	1.3	4.7	4.5	3.6
$\lambda = 100$	5.0	6.5	5.2	5.3	0.2	2.2	3.2	2.2
$\lambda = 200$	5.2	8.5	6.2	5.6	0.02	0.24	2.3	1.9
$\lambda = 300$	5.2	8.7	6.8	6.1	0.05	0.05	1.7	1.4
$\lambda = 400$	5.2	8.7	7.6	6.2	0.01	0	0.9	1.4
$\lambda = 500$	5.2	8.7	7.6	6.2	0.01	0	0.9	1.4
$\lambda = 1000$	5.2	8.7	8.5	7.5	0	0	0	0

Fig. 3 show a comparison between true ERP effect size in human experiment and simulated effect size when $\lambda = 300$. The model prediction aligns well with the real N400 effect size, but it underestimate the size of P600 effect. This is because GPT-2 underestimates the total surprisal of word when it has a syntactic violation.



(a) N400 (Human) (b) N400 (Simulated) (c) P600 (Human) (d) P600 (Simulated)

Figure 3: N400 and P600 effect size from human experiment and from model prediction ($\lambda = 400$).