A noisy channel model of N400 and P600 effects in sentence processing

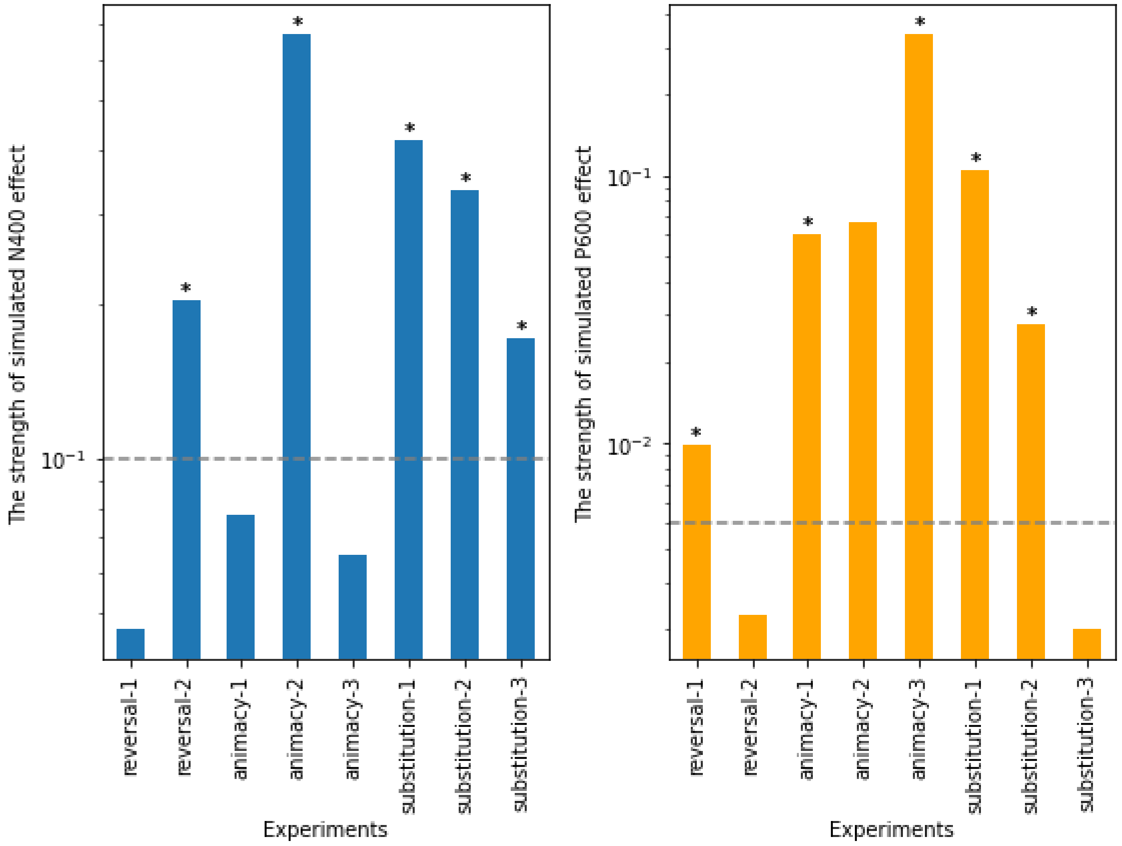
Jiaxuan Li & Allyson Ettinger (University of Chicago)

Introduction: N400 and P600 event-related potential (ERP) components have long been the object of study in psycholinguistics. Traditional accounts have associated N400 effects with semantic violations, and P600 effects with syntactic violations [1,2]. However, this picture is complicated by P600 effects—without N400 effects—in response to animacy [3,4] and thematic-role [5] violations (but only sometimes [6]), as well as biphasic N400/P600 effects for conventional semantic violations [5]. Building on explanations involving interplay of plausibility-driven and syntax-driven interpretations [3,7], we present a computational model that accounts for these complicating observations via a noisy channel modeling framework. Our model assumes early-stage sentence interpretations determined by noisy channel computation (influenced by plausibility), with these early interpretations driving the N400 amplitude. The P600 amplitude reflects reconciliation of the early interpretation with the true (syntax-driven) interpretation, and is modulated by the extent to which early interpretations deviate from the true input. Running this model on original experimental stimuli, we successfully simulate N400 and P600 effects from seven studies in this literature [3-6]. **Method:** We use original stimuli from psycholinguistic experiments featuring semantic / thematic violations, with empirical results varying between N400 effect only, P600 effect only, and biphasic N400/P600 effect (see Table 1). Our use of real experimental stimuli is of note because computational psycholinguistic models often use idealized inputs, while we account for idiosyncratic properties of the real stimuli. To estimate relevant properties of these stimuli (e.g., plausibility, semantic similarity), we draw on outputs of pre-trained models used in natural language processing (NLP). **Noisy channel model:** We implement a noisy channel model to estimate posterior probabilities of potential early interpretations (S_i) given presented input (S_p). These posterior probabilities are based on a) the prior probability of S_i, and b) the likelihood of seeing S_p as a distortion of S_i. For the prior $P(S_i)$, we aim to capture interpretation plausibility, which we approximate via sentence probability estimates from a large neural network pre-trained on word prediction (OpenAI GPT) [8]. We base the likelihood $P(S_p|S_i)$ on the Levenshtein edit distance between S_i and S_p, to capture stronger likelihood of smaller deviations from true input. For each stimulus item S_p, we compute posterior interpretation probabilities for the true input itself, and for one alternative (for anomalous items, a plausible alternative; for control items, an anomalous counterpart). The interpretation with the higher posterior probability is identified as the *early interpretation*. **N400 simulation:** N400 amplitude is approximated by the neural network probability of the target word, given prior context, within the selected early interpretation. **P600 simulation:** To capture reconciliation between interpretations, P600 amplitude is simulated as difference between representations of the early interpretation and the true input, obtained from a neural network pre-trained to detect semantic similarity (fine-tuned DistilBERT) [9]. **Results:** Simulated response amplitudes are averaged by condition, and effects are determined by amplitude differences between critical and control conditions. Results are shown in Fig 1. We see that the model successfully predicts N400 and P600 effects from seven of our eight target experiments. The one failure is a P600 effect appearing for animacy-2 [3]—but we believe that this can be attributed to limitations in the pre-trained neural networks (which show signs of particularly poor estimates on the stimuli in this experiment), rather than to fundamental limitations of our model. **Conclusions:** These results support an account of sentence processing involving early, plausibility-driven interpretation stages (informed by rational inference), reflected in the N400—followed by reconciliation with syntax-driven interpretations, reflected in the P600. Prior work has posited plausibility/syntax interplay [3,7], and other work has linked predictions of noisy channel models to patterns in comprehenders' final interpretations [10,11], and in the P600 [12]. However, to our knowledge this is the first fully-specified computational formalization of plausibility/syntax interplay, the first implemented noisy channel model for simulation of N400 and P600, and the first model of either type to carry out direct prediction of both N400 and P600 components, using real experimental stimuli, across this range of experiments.

Table 1. List of simulated experiments, with experimental manipulations and results.

ID	Manipulation	Violation type	Result	Source
reversal-1	role-reversal	Thematic role	P600	[5]
reversal-2	role-reversal	Thematic role	N400	[6]
animacy-1	Active/passive	Animacy	P600	[3]
animacy-2	Active/passive	Animacy	N400	[3]
animacy-3	Active/passive	Animacy	P600	[4]
substitution-1	word substitution	Lexical meaning	N400 & P600	[5]
substitution-2	word substitution	Lexical meaning	N400 & P600	[5]
substitution-3	word substitution	Lexical meaning	N400	[5]

Fig.1. Simulated N400 (left) and P600 effects (right) across experiments. * represents significant N400/P600 effect in the original human experiment. Dotted line represents a threshold (determined post-hoc) allowing for delineation between presence and absence of effect.



Reference

[1] Kutas & Hillyard (1980). *Biological psychology*.
[2] Hagoort, Brown, & Groothusen (1993). *Language and cognitive processes*.
[3] Kim & Osterhout (2005). *Journal of memory and language*.
[4] Kuperberg, Choi, Cohn, Paczynski, & Jackendoff (2010). *Journal of cognitive neuroscience*.
[5] Chow, Smith, Lau & Phillips (2016). *Language, Cognition and Neuroscience*.
[6] Ehrenhofer, Lau, & Phillips (2020). Forthcoming, (Submitted to Neuropsychologia)
[7] Kuperberg (2007). *Brain research*.
[8] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Preprint.*
[9] Reimers, N., & Gurevych, I. (2019, November). In *EMNLP-IJCNLP.*
[10] Levy (2008, October). In *Proceedings of the 2008 EMNLP*.
[11] Gibson, Bergen & Piantadosi (2013). *Proceedings of the National Academy of Sciences*.
[12] Ryskin, Stearns, Bergen, Eddy, Fedorenko & Gibson (2020). *bioRxiv 2020.02.08.930214*.