A decomposition of surprisal predicts N400 and P600 in language processing

Jiaxuan Li (University of California Irvine) & Richard Futrell (University of California Irvine)
jiaxuan.li@uci.edu

Introduction We propose a new theory of the N400 and P600 based on a decomposition of Surprisal Theory [1, 2] that reflects distinct cognitive mechanisms in language processing. We argue that surprisal can be decomposed into two components: (A) **heuristic surprise** that is correlated with N400, which signals processing difficulty of word given an inferred context; and (B) **structural update** which predicts P600, reflecting the effort of updates to beliefs about previous structure. We validate our theory with (i) quantitative experiments where we tested how well surprisal predicts the summed amplitude of N400 and P600, and (ii) qualitative experiments where we simulate ERP patterns elicited by a variety of linguistic manipulations.

Theory We formalize the idea of a heuristic interpretation in language processing as an inference process in the generative model in Fig. 1. We quantify processing difficulty as the KL divergence between the inferred structure T before and after observing a word W_t (Eq. 1), which yields Surprisal Theory [1, 2]. We observe that we can decompose the predicted processing difficulty into two terms (Eq. 2), where A represents the surprise of a word given the inferred previous structure T_p, and B represents the size of the update to the previous structure. Crucially, the inferred previous structure T_p may involve different words from the veridical context, forming a "heuristic context" resulting from a noisy-channel error correction process [3, 4]. We propose that A and B correspond to the N400 and P600 respectively.

Quantitative validation Our theory makes the following predictions: there is a positive main effect of N400 and P600 on surprisal ("LM surprisal"); N400 is jointly predicted by P600 (with a negative sign) and surprisal (with a positive sign; "LM N400"), and P600 is predicted by N400 (with a negative sign) and surprisal (with a positive sign; "LM P600"). We tested our predictions using a dataset with two experiments (Table 1) [5]. Based on the analysis in [5], we selected four mid-posterior electrodes and averaged ERP amplitudes between 300-500ms as N400, and between 700-900ms as P600. The surprisal of target word is calculated with a large-scale pre-trained neural network (GPT-2 [6]). The results support our predictions (Table 1).

Qualitative validation The main challenge in implementing our theory is to approximate the distribution on inferred past structures T_p (Eq. 3). We used the off-the-shelf T5 grammar corrector, an NLP system that attempts to correct grammatical errors in sentences [7]; the system produces one candidate correction per input sentence, which may be identical to the input sentence. As an approximation, we assume that the distribution on past structures T_p concentrates all its probability mass on this candidate. Then the surprisal given the veridical context and the heuristic surprisal given the past structure is calculated with GPT-2. We simulated experiments on semantic and/or syntactic violation [8], (non)-attractive animacy violation [9] and form/semantic relatedness [10]. The simulated patterns are generally consistent with empirical results (Fig. 2), though with some divergence in [10]—which we attribute to poor T5 approximation of the past structure T_p on the relevant sentences, because T5 does not typically correct lexical errors. Ideally, the distribution on T_p should reflect comprehenders' likely beliefs about the intended past words and structures. In the future, we will estimate the distribution on past structures using a human offline sentence correction task.

Conclusion We propose a surprisal-based theory that predicts N400 and P600. The model provides an information-theoretic model of ERP components grounded on cognitive processes, and sheds light on a fully-specified neurocomputational model of language processing.

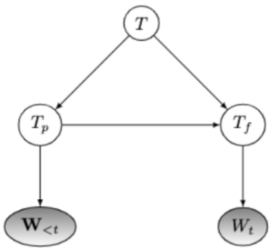
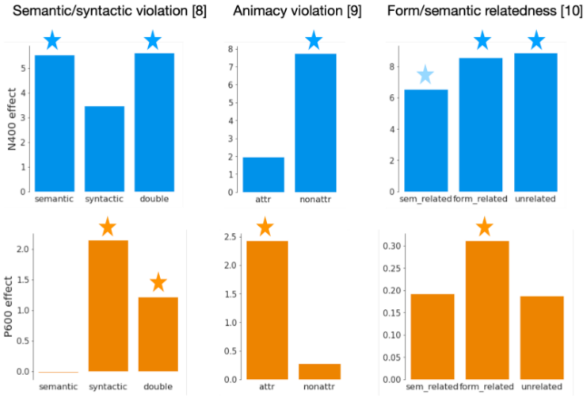


Fig. 1. The comprehender's generative model. Given a sequence of words W_t, comprehenders infer the past structure T_p. When the target word W_t is perceived, comprehenders infer the future structure T_f based on past structure T_p and W_t.


Semantic/syntactic violation [8]

Animacy violation [9]

Form/semantic relatedness [10]

Fig. 2. Simulated N400 (upper) and P600 effect (lower) in Experiment [8]-[9]. The ERP effect is the difference of simulated ERP amplitudes between experimental and control conditions. Blue star ★ represents significant N400 effect in **human** experiment, and orange star ★ represents significant P600 effect. The N400 effect size in sem_related condition of [10] is smaller than in other conditions.

Eq. 1. KL Divergence

$$D \equiv \mathrm{D_{KL}}\left[p\left(T \mid w_t\right) \| p\left(T\right)\right]$$

Eq. 2. Decomposition

$$D = \underbrace{\left\langle -\log p\left(w_t \mid T_p\right)\right\rangle_{p\left(T_p \mid w_{\leq t}\right)}}_{\text{heuristic surprise,}\equiv A} + \underbrace{\left\langle \log \frac{p\left(T_p \mid w_{\leq t}\right)}{p\left(T_p \mid w_{<t}\right)}\right\rangle_{p\left(T_p \mid w_{\leq t}\right)}}_{\text{structural update,}\equiv B}$$

Eq. 3. Bayes Rule

$$p(T_p \mid w_{\leq t}) \propto p(T_p)p(w_{\leq t} \mid T_p)$$

Linear Models

LM surprisal: surprisal ~ N400 + P600 + (1+ N400 + P600 | item) + (1+ N400 + P600 | subject)
LM N400: N400 ~ P600 + surprisal + (1+P600+surprisal | item) + (1+P600+surprisal | subject)
LM P600: P600 ~ N400 + surprisal + (1+N400 + surprisal | item) + (1+P600+surprisal | subject)

Table 1. Quantitative Validation: Experiment stimuli and statistical analysis of model predictions. Numbers are t-values. p < 0.005***.

Exp	Manipulation	Empirical effect	LM surprisal	LM N400	LM P600
Exp1	Substitution	biphasic	N400: t = 3.33*** P600: t = 2.57***	Surprisal: t = 5.00*** P600: t = -33.07***	Surprisal: t = 4.95*** N400: t = -27.09***
	Reversal	P600			
Exp2	Substitution	biphasic	N400: t = 2.00*** P600: t = 1.69***	Surprisal: t = 2.66*** P600: t = -30.42***	Surprisal: t = 2.00*** N400: t = -24.18***
	Swap	N400			

References: [1] Hale et al., 2001. NAACL. [2] Levy, 2008. *Cognition*. [3] Levy et al., 2009. *PNAS*. [4] Ryskin et al., 2021. *Neuropsychologia*. [5] Chow et al., 2016. *Language, Cognition and Neuroscience*. [6] Radford et al., 2019. *Arkiv preprint*. [7] Raffel et al., 2019. *JMLR*. [8] Ainsworth-Darnell et al., 1998. *Journal of Memery and Language (JML)*. [9] Kim & Osterhout, 2005. *JML*. [10] Ito et al., 2016. *JML*.