A decomposition of surprisal tracks the N400 and P600 brain potentials

Jiaxuan Li **Richard Futrell**
Department of Language Science
University of California Irvine
Irvine, CA 92617
{jiaxuan.li,rfutrell}@uci.edu

Abstract

The functional interpretation of language-related ERP components has been a central debate in psycholinguistics for decades. 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 the 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 four experiments. Our theory is in principle compatible with traditional cognitive theories assuming the existence of a 'good-enough' heuristic interpretation, but with a precise information-theoretic formulation.

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 (Kutas & Hillyard, 1980; Hagoort et al., 1993; Hoeks et al., 2004; Kim & Osterhout, 2005; Van Herten et al., 2005; Van Petten & Luka, 2012; Kuperberg, 2007, 2016; Van Petten & Luka, 2012).

Traditionally, the N400 and P600 ERP components have been linked to semantic and syntactic anomalies in sentences, and thus taken to indicate some degree of modularity with respect to syntactic and semantic processing (Kutas & Hillyard, 1980; Hagoort et al., 1993). However, a number of studies have found P600 effects in response to semantic violations, with or without a corresponding N400 (Kim & Osterhout, 2005; Kuperberg, 2007; Chow et al., 2016; Brouwer et al., 2012; Ryskin et al., 2021; Ito et al., 2016; Van Petten & Luka, 2012). In response to these data, 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 (Van Herten et al., 2005, 2006; Kuperberg, 2016; Brouwer et al., 2012; Ferreira et al., 2002; Ferreira & Stacey, 2000; Ferreira & Patson, 2007), along with an error monitoring process. In such theories, the N400 reflects how well the heuristic interpretation is semantically well-formed, whereas P600 reflect deviance between heuristic and literal interpretations. However, none of the theories have succeeded in explaining the full range of empirical results (see Brouwer et al., 2012, for detailed discussion).

Prior work has suggested to formalize this heuristic interpretation/error monitoring process within a noisy-channel framework. The noisy-channel model posits that comprehenders rationally infer a probabilistic distribution on the intended meaning given the received input while taking into account the fact that the input may contain errors (generically termed "noise"). Rational inference in this setting involves a trade-off between the prior of how likely the intended message is conveyed, and the likelihood of how likely the intended message is to be distorted into the perceived message. Prior work has established that there is a reduced N400 and a larger P600 when a correct sentence can be recovered from original sentence with a semantic error (Gibson et al., 2013; Ryskin et al., 2021). Further, computational models have been developed where the amplitudes of N400 and P600 signal reflect distinct aspects of language processing over the inferred heuristic interpretation (Li & Ettinger, 2023). However, these studies have subjective decisions about possible candidates for heuristic interpretations, making it difficult to scale up and account for other studies. In addition, these studies do not have a principled information-theoretic quantification of cognitive effort, and are not integrated with more general computational neuroscientific models of other cognitive processes (Ortega & Braun, 2013; Gershman, 2020; Futrell & Hahn, 2022; Zénon et al., 2019).

We propose an information-theoretic computational-level model of the N400 and P600 ERP components in language processing. Our model formalizes the noisy-channel framework described above and situates it as a generalization of Surprisal Theory, a probabilistic model of online language comprehension (Hale, 2001; Levy, 2008). We argue that surprisal can be decomposed into two parts: (A) the **heuristic**

surprise of the current word within an inferred heuristic distribution of utterances, predicting the magnitude of the N400 signal, and (ii) a **discrepancy signal** reflecting the difference between the surprisal of the veridical input and the surprisal of the heuristically-inferred utterance, predicting the magnitude of the P600 signal. We run qualitative simulations over previously-reported data from three experiments, featuring semantic and syntactic violations, event structure violations, semantic relatedness priming. We further perform quantitative analyses on one additional experiment, where we show that our quantities for heuristic surprise and discrepancy signal track the N400 and P600 components respectively. Our model successfully explains a wide range of ERP patterns and integrates multiple strands of psycholinguistic research into a quantitative model. 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. Furthermore, we leverage recent computational models from the field of natural language processing to implement our theory, which allows us to formalize the probabilistic process of language comprehension with statistical properties of real experimental inputs.

Model

Our model builds on Surprisal Theory, an empirically successful theory of behavioral signatures of language comprehension such as reading time (Hale, 2001; Levy, 2008; Frank & Bod, 2011; Smith & Levy, 2013; Wilcox et al., 2020), which is in line with recent computational neuroscientific proposals to quantify cognitive effort information-theoretically (Zénon et al., 2019). 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. Fig. 1 shows an overview of model architecture.

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 2. 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, or there may be errors in perception, and comprehenders should be able to correct for these factors. We formalize this idea by introducing random variables for **heuristic words**

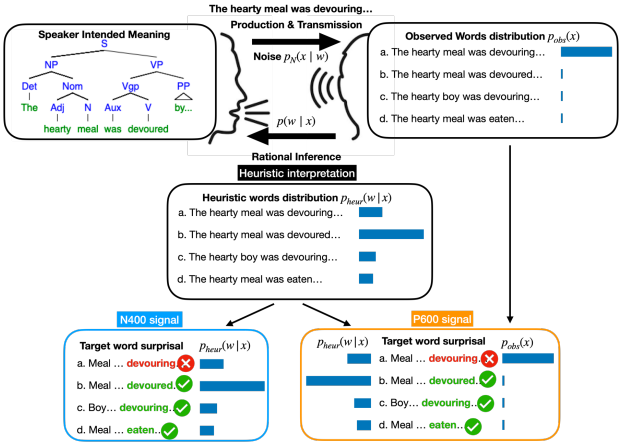


Figure 1: Overview of model architecture. A speaker intends to convey "The hearty meal was devoured", but the comprehender's observed input is "The hearty meal was devouring".

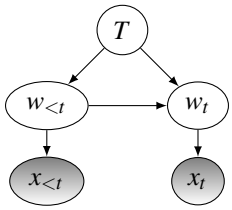


Figure 2: The comprehender's generative model. T is the speaker's intended structure. At time t, the structure T contains words $w_{<t}$ and (the past context) and w_t (current word). The comprehender observes a noisy form of the past context $x_{<t}$ and the current word x_t.

$w_{<t}$ and w_t, corresponding to the values of past words and the current word within the speaker's intended 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.

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 2, the surprisal S_t $(-\ln p(x_t \mid x_{<t}))$ can be partitioned into two parts,[1] 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**:

$$S_t = \underbrace{\mathbb{E}\left[-\ln p(w_t \mid w_{<t})\right]}_{\text{heuristic surprise},=A} + \underbrace{\mathbb{E}\left[\ln \frac{p(w_t \mid w_{<t})}{p(x_t \mid x_{<t})}\right]}_{\text{discrepancy signal},=B}, \tag{2}$$

where the expectations are with respect to the probability distribution $p(w_{\leq t} \mid x_{\leq t})$. The heuristic surprise is an upper

[1] Our decomposition of surprisal is distinct from the decomposition into lexical and syntactic surprisal proposed by Roark et al. (2009).

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

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

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 assume a noise model based on the relationship between heuristic words w and input words x in terms of both form and meaning:

$$p_N(x \mid w) \propto \exp\{-\lambda [d(x,w) + \gamma s(x,w)]\}, \tag{5}$$

where $d(x,w)$ is a phonological or orthographic distance from w to x and $s(x,w)$ is the semantic similarity of x and w. These two factors reflect common sources of errors in speech (Dell & Reich, 1981). The scalar free parameters are λ, representing the overall (inverse) rate of errors, and γ, representing the relative importance of form-based and meaning-based factors. The value of γ is held constant across experiments simulated below, while the overall inverse noise rate λ varies from experiment to experiment, reflecting differences in experimental task, number of implausible stimuli, etc.

In order to compute Eq. 2, we need a way to average over all the possible heuristic word strings w given the input word strings x. This averaging is technically difficult as there are in principle an infinite number of possible w underlying any given input x. Therefore, we compute averages approximately, limiting the support of w to only a subset of likely candidates generated by prompting GPT-3 (specifically `text-davinci-002` Brown et al., 2020). For each experimental stimulus, we input prompts with instructions and four examples on how to correct sentences to recover its intended meaning[2], and ask GPT-3 to generate one best correction. We generate ten corrections for each sentence when the model parameter is encouraged to generate different corrections (GPT-3 model temperature: 0.95).

Once a set of candidate heuristic word strings w is generated in this way, we calculate the likelihood of each string by following Eq. 5 applied independently to the individual words in the string. For the form-based distance $d(x,w)$ we use orthographic Levenshtein edit distance; for the semantic distance $s(x,w)$ we use cosine distance calculated using GPT-2 embeddings (Radford et al., 2019). For the prior $p(w)$, we calculate contextualized probability at target word position using the GPT-2 language model.

Empirical Validation

Datasets

We selected four ERP studies as our dataset. The studies include manipulations featuring a variety of semantic and syntactic violations. Table 1 shows a list of conditions with sample stimuli and empirical ERP patterns across experiments. The ERP effects in the experimental conditions are all calculated in terms of differences to the ERP signal in the control condition. We conducted qualitative analysis on the first three experiments and quantitative analysis for the last one.

The first study (hereby *AD-98*) (Ainsworth-Darnell et al., 1998) includes four conditions: one control condition (*Control*); one condition with violation of semantic content (*Semantic*); one with syntactic manipulation that shows a P600 effect (*Syntactic*); and one with both semantic and syntactic violation (*Double*). The conditions with semantic violation (*Semantic* and *Double*) elicited an N400 effect relative the *Control* condition, whereas the conditions with syntactic violation (*Syntactic* and *Double*) triggered a P600 effect.

In the second study (hereby *Kim-05*), from Kim & Osterhout (2005), there are three conditions: *Attractive*, *Non-attractive*, *Control*. In experimental conditions the animacy of the subject is violated. Additionally, the semantic association between subject and target verb is manipulated such that subject and verb could form a plausible event (in *Attractive* condition) or not (in *Non-attractive*). While *Attractive* condition elicited a greater P600 response compared to the control condition, the *Non-attractive* condition elicited a greater N400.

The third study (hereby *Ito-16*) from Ito et al. (2016) includes four conditions. The three experimental conditions all change one target word in the *Control* condition into a semantically implausible one. In *Semantic-related* condition, the semantic violation is semantically related to the target in *Control* condition. In *Form-related* condition, the semantic error shares orthographic form with the *Control* target. In *Unrelated* condition, the violation is not related to the target in the *Control* condition. All three experimental manipulations triggered N400 effects. In addition, the size of N400 effect to semantically related violation was reduced, and form-related violation triggered a P600 effect as well.

The last experiment (hereby *Ryskin-21*) from Ryskin et al. (2021) has four conditions, one with a semantic violation (*Semantic*), one with a syntactic violation (*Syntactic*), one seman-

[2]Prompt: The final word in each of the following sentences is wrong: someone typed the wrong word. Please type in a different word, the one that was probably intended. **Input**: The hearty meal was devouring. **Correction**: The hearty meal was devoured. **Input**: The hearty meal was devoured. **Correction**: The hearty meal was devoured. **Input**: Mary went to the library to borrow a hook. **Correction**: Mary went to the library to borrow a book. **Input**: Mary went to the library to borrow a plant. **Correction**: Mary went to the library to borrow a plant. **Input**: *Experimental Sentence* **Correction**:

Experiment	Condition	Sentence	Empirical ERP
AD-98	Syntactic	The victims reported robbery markets.	P600
	Semantic	The victims reported robbery to water.	N400
	Double	The victims reported robbery water.	Biphasic
	Control	The victims reported robbery to markets.	NA
Kim-05	Attractive	The hearty meal was devouring...	P600
	Non-attractive	The dusty tabletop was devouring...	N400
	Control	The hearty meal was devoured...	NA
Ito-16	Semantic-related	The student is going to the library to borrow a page...	Reduced N400
	Form-related	The student is going to the library to borrow a hook...	Biphasic
	Unrelated	The student is going to the library to borrow a sofa...	N400
	Control	The student is going to the library to borrow a book...	NA
Ryskin-21	Semantic	The storyteller could turn any incident into an amusing hearse.	N400
	Syntactic	The storyteller could turn any incident into an amusing anecdotes.	P600
	Recoverable	The storyteller could turn any incident into an amusing antidote.	Biphasic
	Control	The storyteller could turn any incident into an amusing anecdote.	NA

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

tic critical condition (*Recoverable*) with a semantic violation which could be attributed to noise, and a control sentence without any error (*Control*). In the N400 time window, there is a significant N400 effect in *Semantic* and *Recoverable* conditions, where the N400 effect in *Recoverable* condition is reduced. In the P600 time window, there is a significant P600 effect in *Syntactic* and a smaller but significant P600 effect in *Recoverable* condition (see Fig. 4b).

Free parameters

We have two free parameters in the model: an experiment-dependent parameter λ that reflects variations in experimental set-up, and an experiment-independent parameter γ that accounts for relative importance of semantic and form similarity between heuristic and true interpretations.

We first set γ to be 1 and explored the effect of λ with a grid search from 100 to 600, with a step size of 10, and with two marginal conditions ($\lambda = 0$ and $\lambda = 1000$). When $\lambda = 0$, the heuristic surprise is the surprisal of the most predictable word given the context, regardless of the true target received. As λ increases, it becomes more difficult to do error correction, resulting in an increased heuristic surprise and a decreased discrepancy signal. For each experiment, we selected the λ based on visual inspection of the simulated ERP pattern. Table 2 shows a list of λ values used in simulating the presented results. This parameter reflects experiment-specific tendency to arrive at a heuristic interpretation that is different from literal input. Many factors have been reported to affect the size of N400/P600 effect, including the proportion of plausible/implausible sentences in the stimuli, the nature of task demands and presentation latency (Gunter et al., 1997; Hahne & Friederici, 1999; Zwaan & Radvansky, 1998; Chow et al., 2018).

Next, having set λ to the values, we performed a grid search of γ from 0 to 1 with a step size of 0.1. We did another round of visual inspection of the simulated ERP patterns and selected $\gamma = 0.8$ for all experiments.

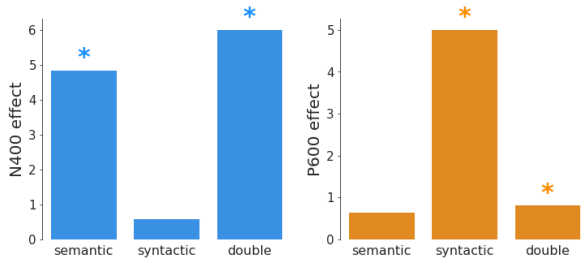
	AD-98	Kim-05	Ito-16	Ryskin-21
λ	180	150	590	320

Table 2: λ values across experiments

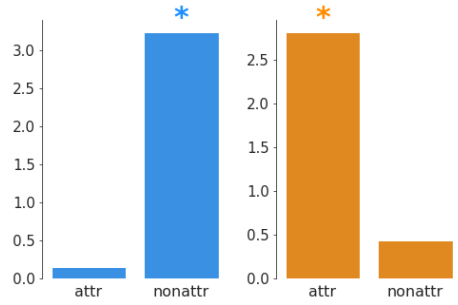
Qualitative results

Fig. 3a shows the simulated ERP patterns in AD-98 (Ainsworth-Darnell et al., 1998). As expected, our model successfully simulated a larger N400 effect in conditions where there is a semantic violation (*Semantic* and *Double* conditions), and a larger P600 effect in *Syntactic* and *Double* conditions. However, we want to acknowledge that model might have underestimated the size of P600 effect in *Double* condition, due to our generation of alternatives with GPT-3. In our prompt design, GPT-3 is only incentivized to correct sentence if the correction could form a plausible interpretation with a small number of edits, but unlikely to dissociate semantic and syntactic violations and correct syntactic violations. For example, GPT-3 corrected the sentence with double violation"*The victim reported the robbery markets*" into "*The victim reported the robbery masterminds*" instead of "*The victim reported the robbery to markets*". Therefore, our heuristic words *w* generated by GPT-3 tend to remain the same as original input, and does not reflect the distribution of heuristic interpretations in humans.

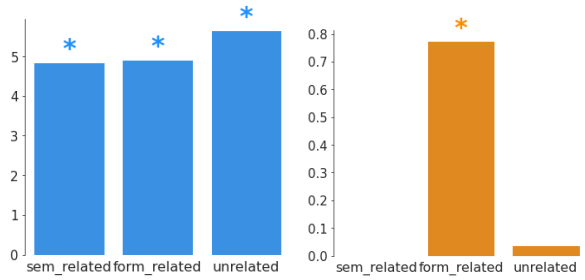
Fig. 3b shows N400 and P600 effects in Kim-05 (Kim & Osterhout, 2005). Consistent with human experimental results, the model simulated a greater N400 amplitude to *Non-*



(a) AD-98 (Model)



(b) Kim-05 (Model)



(c) Ito-16 (Model)

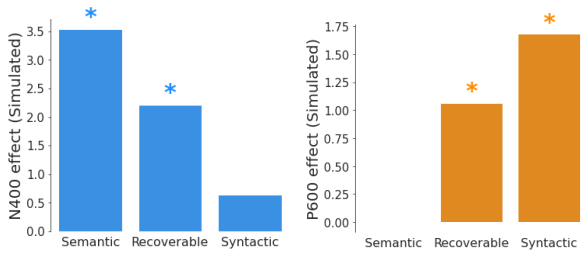
Figure 3: N400 (left) and P600 (right) amplitudes from model simulation in (a) AD-98 (upper), (b) Kim-5 (middle) and (c) Ito-16 (bottom). Blue * indicates a significant N400 effect in the human experiment, and orange * indicates a significant P600 effect

attractive animacy violation, and a greater P600 response in *Attractive* condition, relative to *Control* condition. The success of the model relies on the fact that a larger proportion of sentences in *Attractive* condition have been error corrected to more plausible heuristic alternatives than in *Non-attractive* condition.

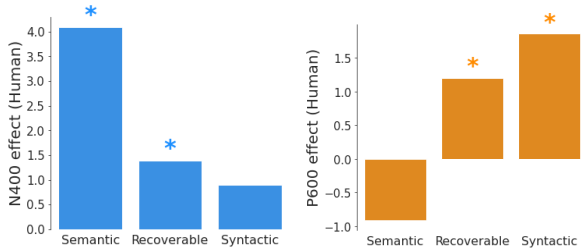
In Ito-16 (Ito et al., 2016) (see Fig. 3c), the model correctly predicts an N400 effect in *Form-related*, *Semantic-related* and *Unrelated* conditions, relative to *Control* conditions, and a P600 effect in *Form-related* condition. Importantly, consistent with human ERP patterns, the N400 effect is reduced for *Semantic-related* and *Form-related* conditions, where N400 effect to *Semantic-related* conditions being smallest. Our results provide convincing and distinct explanation for observed ERP effects in *Semantic-related* and *Form-related* conditions. In the *Semantic-related* condition, the semantic similarity term $s(x,w)$ prevents semantic-related words to be corrected into plausible targets. The N400 reduction for semantic-related words is induced by associative pre-activation the semantic meaning, as reflected in the reduction of veridical surprisal of target word. In contrast, in the *Form-related* condition, the phonological distance term $d(x,w)$ encourages form-related errors to be corrected. The N400 reduction in the *Form-related* condition is a result of assigning a more plausible interpretation to errors similar to the true prediction in the surface form.

Quantitative Validation

We performed a quantitative analysis on the Ryskin-21 dataset (Ryskin et al., 2021). Fig. 4a shows the model simulated ERP effects across conditions, and Fig. 4b shows empirical ERP effects in human experiments. The model successfully simulates an N400 for semantic violations, a P600 effect for syntactic violations, and a biphasic effect for recoverable semantic violations. More importantly, the model tracks the order of magnitude of N400 and P600 effects, with a greater N400 effect in *Semantic* than *Syntactic* condition, and a greater P600 effect in *Syntactic* than *Recoverable* condition.



(a) Model



(b) Human

Figure 4: N400 (left) and P600 (right) amplitudes from model simulation (upper) and from human ERP experiments (lower) in Ryskin-21. Blue * indicates a significant N400 effect in the human experiment, and orange * indicates a significant P600 effect.

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 (Barr et al., 2013). We use heuristic surprise as a single predictor to predict ERP amplitude in the N400 time window, and discrepancy signal to predict ERP amplitude in the P600 time window. We additionally include two models where veridical surprisal is used

to predict N400 and P600 amplitude as a comparison. The models have by-subject and by-item random intercept and slopes. The operationalization of N400 and P600 amplitude is based on analysis in original studies (Ryskin et al., 2021). We selected averaged ERP amplitudes from eight central-parietal electrodes, and averaged across 300-500ms N400 time window and 600-800 P600 time window. The surprisal of target word is calculated with GPT-2.

Table 3 shows the results. We find a significant main effect of heuristic surprise on N400 amplitude ($t = -6.77, p < .01$), and a significant main effect of discrepancy signal on P600 amplitude ($t = 2.88, p < .01$). In comparison, we find no significant effect of veridical surprisal on P600 ($t = 0.20, p = 0.20$), suggesting that our proposed decomposition of surprisal provides a better fit of the overall ERP components than veridical surprisal alone. This finding is not contradictory to the fact that veridical surprisal calculated with large-scale pre-trained language models can predict N400 effects in many cases (Frank et al., 2015; Michaelov et al., 2021; Michaelov & Bergen, 2020): In many cases, heuristic surprisal and veridical surprisal are similar.

N400		P600	
heuristic surprise	verid. surprisal	discrepancy signal	verid. surprisal
-0.68 (-6.77**)	-0.59 (-5.78***)	0.59 (2.88**)	0.02 (0.20)

Table 3: The effects of veridical surprisal S_t (the surprisal of the veridical input x_t given veridical context $x_{<t}$), heuristic surprise A (the average surprisal of the heuristic inputs w_t given heuristic contexts $w_{<t}$), and discrepancy signal B on ERP amplitudes in the N400 and P600 time range in the experiment from (Ryskin et al., 2021). 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.

Discussion

The functional interpretation of N400 and P600 has been a central debate in psycholinguistics. We presented a neurocomputational 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 correspond to N400 and P600 respectively. The two measures have a clear cognitive interpretation. The heuristic surprise signals the processing difficulty associated with the information provided by the heuristic word, and the discrepancy signal represents the effort of error identification and correction. We derive the distribution on heuristic interpretations via noisy-channel inference, and implement it with large-scale language models. Our model can not only simulate ERP patterns featuring a variety of linguistic manipulation, but also track the amplitude of N400 and P600 at the item level.

What distinctive characteristics contributes to the success of our model? How is our model related to a broader ERP landscape? Our model assumes a heuristic interpretation stage in language processing. Both behavioral and neural studies have suggested that comprehenders might use shallow and heuristic cues to form a plausibility-based "good-enough" interpretation (Ferreira et al., 2002). However, the nature of the heuristic interpretation and the strategies use of linguistic cues has been under-specified. The noisy-channel model for heuristic interpretations 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 our implementation of heuristic interpretation provides a good fit to empirical data, we want to acknowledge that this model is on a computational level, and further work could be done on the precise algorithmic nature of the heuristic interpretation generation process.

Our model interprets the N400 signal as an index of contextualized word processing difficulty over heuristic interpretations. Theoretical views of N400 have diverged in whether N400 indexes pre-activation of upcoming linguistic inputs, or integration effort of the target word given previous representation, or a combinatory process of both (see Kutas & Federmeier, 2011, for more discussion). Our use of contextualized word probability metrics is compatible with all these theoretical perspectives, and also consistent with other computational work on N400 (Rabovsky et al., 2018; Brouwer et al., 2017, 2021; Fitz & Chang, 2019; Li & Ettinger, 2023; Michaelov & Bergen, 2020; Michalon & Baggio, 2019). More importantly, our model argues that N400 is associated with cognitive processes on a heuristic representation that could differ from the true literal meaning of the input. This allows our model to provide good predictions of emerging experimental evidence on N400 blindness to semantic anomalies with a plausible alternative (Chow et al., 2016; Kim & Osterhout, 2005; Kuperberg, 2016; Van Herten et al., 2006).

Our model hypothesizes that P600 indexes the effort of conflict resolution or discourse update between inferred heuristic interpretation and the true literal interpretation, in line with multiple strands of theoretical frameworks (Kim & Osterhout, 2005; Van Herten et al., 2005; Leckey & Federmeier, 2020; Van Petten & Luka, 2012). Our model argues that syntactic and semantic P600 operate under the same cognitive mechanisms of belief update with respect to previous interpretations.

Our model links neural signatures to behavioral measures via Surprisal Theory. Our model predicts that N400 and P600 index different aspects of processing difficulties, and the summed ERP amplitudes correspond to word processing difficulties indexed by other behavioral measures such as reading time. The work calls for co-registration of brain and behavioral experiments.

References

Ainsworth-Darnell, K., Shulman, H. G., & Boland, J. E. (1998). Dissociating brain responses to syntactic and semantic anomalies: Evidence from event-related potentials. *Journal of Memory and Language*, *38*(1), 112–130.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. (2017). A neurocomputational model of the N400 and the P600 in language processing. *Cognitive science*, *41*, 1318–1352.

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*.

Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about semantic illusions: rethinking the functional role of the P600 in language comprehension. *Brain research*, *1446*, 127–143.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... others (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Chow, W.-Y., Lau, E., Wang, S., & Phillips, C. (2018). Wait a second! Delayed impact of argument roles on on-line verb prediction. *Language, Cognition and Neuroscience*, *33*(7), 803–828.

Chow, W.-Y., Smith, C., Lau, E., & Phillips, C. (2016). A "bag-of-arguments" mechanism for initial verb predictions. *Language, Cognition and Neuroscience*, *31*(5), 577–596.

Dell, G. S., & Reich, P. A. (1981). Stages in sentence production: An analysis of speech error data. *Journal of verbal learning and verbal behavior*, *20*(6), 611–629.

Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current directions in psychological science*, *11*(1), 11–15.

Ferreira, F., & Patson, N. D. (2007). The 'good enough'approach to language comprehension. *Language and Linguistics Compass*, *1*(1-2), 71–83.

Ferreira, F., & Stacey, J. (2000). The misinterpretation of passive sentences. *Manuscript submitted for publication*. (publisher: Citeseer)

Fitz, H., & Chang, F. (2019). Language ERPs reflect learning through prediction error propagation. *Cognitive Psychology*, *111*, 15–52.

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, *22*(6), 829–834.

Frank, S. L., Otten, L., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and language*, *140*, 1–11.

Futrell, R., & Hahn, M. (2022). Information theory as a bridge between language function and language form. *Frontiers in Communication*, *7*, 657725.

Gershman, S. J. (2020). Origin of perseveration in the trade-off between reward and complexity. *bioRxiv*.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056.

Gunter, T. C., Stowe, L. A., & Mulder, G. (1997). When syntax meets semantics. *Psychophysiology*, *34*(6), 660–676.

Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and cognitive processes*, *8*(4), 439–483.

Hahne, A., & Friederici, A. D. (1999). Electrophysiological evidence for two steps in syntactic analysis: Early automatic and late controlled processes. *Journal of cognitive neuroscience*, *11*(2), 194–205.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics.*

Hoeks, J. C., Stowe, L. A., & Doedens, G. (2004). Seeing words in context: the interaction of lexical and sentence level information during reading. *Cognitive brain research*, *19*(1), 59–73.

Ito, A., Corley, M., Pickering, M. J., Martin, A. E., & Nieuwland, M. S. (2016). Predicting form and meaning: Evidence from brain potentials. *Journal of Memory and Language*, *86*, 157–171.

Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of memory and language*, *52*(2), 205–225.

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain research*, *1146*, 23–49.

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

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (ERP). *Annual review of psychology*, *62*, 621–647.

Kutas, M., & Hillyard, S. A. (1980). Event-related brain potentials to semantically inappropriate and surprisingly large words. *Biological psychology*, *11*(2), 99–116.

Leckey, M., & Federmeier, K. D. (2020). The p3b and p600 (s): Positive contributions to language comprehension. *Psychophysiology*, *57*(7), e13351.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Li, J., & Ettinger, A. (2023). Heuristic interpretation as rational inference: A computational model of the n400 and p600 in language processing. *Cognition*, *233*, 105359.

Michaelov, J., & Bergen, B. (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).

Michaelov, J., Coulson, S., & Bergen, B. (2021). *So cloze yet so far: N400 amplitude is better predicted by distributional information than human predictability judgements.*

Michalon, O., & Baggio, G. (2019). Meaning-driven syntactic predictions in a parallel processing architecture: Theory and algorithmic modeling of erp effects. *Neuropsychologia*, *131*, 171–183.

Ortega, P. A., & Braun, D. A. (2013). Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A: Mathematical, Physical, and Engineering Sciences*, *469*, 20120683.

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the n400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, *2*(9), 693–705.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 conference on empirical methods in natural language processing: Volume 1-volume 1* (pp. 324–333).

Ryskin, R., Stearns, L., Bergen, L., Eddy, M., Fedorenko, E., & Gibson, E. (2021). An ERP index of real-time error correction within a noisy-channel framework of human communication. *Neuropsychologia*, *158*, 107855.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Van Herten, M., Chwilla, D. J., & Kolk, H. H. (2006). When heuristics clash with parsing routines: ERP evidence for conflict monitoring in sentence perception. *Journal of cognitive neuroscience*, *18*(7), 1181–1197.

Van Herten, M., Kolk, H. H., & Chwilla, D. J. (2005). An ERP study of P600 effects elicited by semantic anomalies. *Cognitive brain research*, *22*(2), 241–255.

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176–190.

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. In *Proceedings for the 42nd annual meeting of the cognitive science society* (pp. 1707–1713).

Zénon, A., Solopchuk, O., & Pezzulo, G. (2019). An information-theoretic perspective on the costs of cognition. *Neuropsychologia*, *123*, 5–18.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological bulletin*, *123*(2), 162.