Cognition 233 (2023) 105359

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original articles

Heuristic interpretation as rational inference: A computational model of the N400 and P600 in language processing

Jiaxuan Li [a,*], Allyson Ettinger [b]

[a] Department of Language Science, University of California Irvine, USA
[b] Department of Linguistics, University of Chicago, USA

ARTICLE INFO

Dataset link: https://github.com/goldengua/Cognition-Noisy-channel-ERP

Keywords:
Psycholinguistics
Semantic P600
N400
Noisy channel model
Computational psycholinguistics
Semantic anomaly
ERPs

ABSTRACT

Much inquiry in psycholinguistics has focused on evidence from the N400 and P600 components of the event-related potential (ERP) signal—and a central theoretical challenge in this area is accounting for the so-called “semantic P600”, which involves unexpected patterns in these components relative to traditional theories of the underlying mechanisms. In this paper we present a computational model of the language processing mechanisms underlying these ERP components, which builds on existing psycholinguistic theories in positing a *heuristic interpretation* stage of processing, but which deviates from existing theories in formulating this heuristic interpretation process as probabilistic selection via a noisy channel model, and in quantifying and accounting for fine-grained variation in statistical and representational properties of individual stimuli. Our model successfully simulates N400 and P600 patterns from eight psycholinguistic experiments, reflecting the full range of N400-only, P600-only, and biphasic N400-P600 effects, and its behaviors shed light on a number of key patterns that have presented challenges for existing theories. The model’s success indicates that a strong account for the processing mechanisms underlying these effects is one in which language comprehension involves a probabilistic heuristic interpretation stage resembling a noisy channel process, feeding into subsequent processes that assess target word fit and reconcile between heuristic and literal interpretations. The model’s success also indicates that these mechanisms are critically sensitive to statistical variation in individual stimuli, and that modeling the effects of this variation is essential to account for the full range of observed effects in language processing.

1. Introduction

The phenomenon known as the “semantic P600” has presented a continuing point of mystery in psycholinguistics. This phenomenon is detected in association with measurement of event-related potentials (ERPs), and involves unexpected behaviors of the N400 and P600 components of the ERP signal. The N400 component (a negative-going deflection peaking at around 300–500 ms after the onset of the stimulus) and the P600 ERP component (a post-N400 positive inflection roughly in the 600–1000 ms time window) have historically been connected with semantic and syntactic violations, respectively (Brown & Hagoort, 1993; Hagoort, Brown, & Groothusen, 1993; Kutas & Hillyard, 1980; Osterhout & Holcomb, 1992). However, this traditional mapping has been challenged by a number of studies finding P600 effects, instead of N400 effects, in response to semantic violations such as role-reversal and animacy violations. This phenomenon has often been referred to as the “semantic P600 effect” (Erickson & Mattson, 1981; Hoeks, Stowe, & Doedens, 2004; Kim & Osterhout, 2005; Kuperberg, 2007, 2016; Nieuwland & Van Berkum, 2005; Van Herten, Kolk, & Chwilla, 2005; Van Petten & Luka, 2012).

A variety of theories have been proposed to account for this semantic P600 phenomenon. While these theories differ in the precise nature of hypothesized mechanisms, most have in common the involvement of what we will refer to as a *heuristic* interpretation—an interpretation of the input that is in play during processing and that may differ from the literal input. This interpretation is typically believed to be formed on the basis of a subset of information in the input—such as the most plausible combinations of words that have been presented, regardless of actual syntax (Bornkessel-Schlesewsky & Schlesewsky, 2008; Hoeks et al., 2004; Kim & Osterhout, 2005; Kolk, Chwilla, Van Herten, & Oor, 2003; Kuperberg, 2016; Michalon & Baggio, 2019; Van Herten et al., 2005). Theories along this line are typically verbally formulated, and often focus on the details of the different processing streams that may compete to influence interpretations. While each of these theories gives a compelling account for many of the observed experimental results,

[*] Correspondence to: Department of Language Science, University of California Irvine, Irvine, CA 92617, USA.
E-mail addresses: jiaxuan.li@uci.edu (J. Li), aettinger@uchicago.edu (A. Ettinger).

https://doi.org/10.1016/j.cognition.2022.105359
Received 13 April 2022; Received in revised form 28 October 2022; Accepted 9 December 2022
Available online 20 December 2022

for all of these theories there are certain outstanding challenges that remain to be addressed. In particular, existing theories tend not to account for observation of different outcomes under the same manipulation, or instances in which both N400 and P600 effects are observed for a given manipulation (biphasic effects).

In this paper we present a computational model that builds on these existing theories with two key mechanistic additions. Our first addition is that we formalize the heuristic interpretation process as a probabilistic selection between candidate interpretations. We model this selection process via a noisy channel computation, which balances a notion of interpretation plausibility against extent of deviation from the literal input string. This probabilistic process can be conceptualized as an "error-correction" decision, which determines whether the heuristic interpretation of an input will remain true to the literal input, or reflect a minimally-different alternative. After selection of the heuristic interpretation, we then simulate N400 amplitude based on target word probability within the selected heuristic interpretation, and P600 amplitude based on semantic divergence between the heuristic and literal interpretations. This approach builds on prior literature applying noisy channel models for psycholinguistics (Gibson, Bergen, & Piantadosi, 2013; Levy, 2008; Ryskin, Stearns, Bergen, Eddy, Fedorenko, & Gibson, 2021), including work showing a relationship of N400 and P600 amplitudes to probability of error correction (Ryskin et al., 2021).

Our second addition is that we quantify key model variables (plausibility, word probability, semantic divergence) via large pre-trained neural networks, allowing for sensitivity to idiosyncratic statistical variation between stimulus items. Using the real experimental stimuli from the psycholinguistic studies that we simulate, we show that this model is able to overcome both of the above challenges for existing theories, successfully simulating ERP patterns to eight studies with manipulations including role reversals (Chow, Smith, Lau, & Phillips, 2016; Ehrenhofer, Lau, & Phillips, in press), animacy violations (Kim & Osterhout, 2005; Kuperberg, Choi, Cohn, Paczynski, & Jackendoff, 2010), word substitutions (Chow et al., 2016) and syntactic violation (Ainsworth-Darnell, Shulman, & Boland, 1998). One additional simulation shows divergence from human patterns, likely as a result of imperfect estimates from our neural networks, which we discuss in detail.

Our model shows that if heuristic interpretation theories are formulated probabilistically as we do here, and if they consider fine-grained statistical variation among individual stimuli, they can account for a wide range of N400/P600 results, including key outstanding challenges related to the semantic P600. One important insight that comes out of our analysis of the simulation results is that revisions involved in the heuristic interpretation process have valuable explanatory power for these phenomena, but they make up only part of a larger story, in which fine-grained variation in probabilistic and representational stimulus properties also exert important influence on the ultimate effects generated by the posited mechanisms. On the whole, our model lays out a promising account of the processing mechanisms that give rise to observed ERP patterns in these experiments, building on existing theory while strengthening the detail and explanatory power of previous accounts.

2. Background

2.1. Sentence processing models

Observations surrounding the semantic P600 have given rise to a number of related theories involving multiple interacting streams of processing. These theories share the notion of a semantic/plausibility stream, which has the flexibility to process the sentence as a more plausible alternative. Kim and Osterhout (2005) refer to this as Semantic Attraction: if sufficiently semantically attractive, an alternative interpretation can be chosen over the literal input meaning. The Monitoring Theory (Kolk et al., 2003; Van Herten, Chwilla, & Kolk, 2006; Van Herten et al., 2005; Van Schijndel & Linzen, 2021) has a plausibility stream that can produce a temporary heuristic interpretation based on world knowledge, word meaning and surface word order. Continued Combinatory Analysis (CCA) (Kuperberg, 2007) and the extended Argument Dependency Model (eADM) (Bornkessel-Schlesewsky & Schlesewsky, 2008) include a semantic stream plus an additional Thematic stream. The means of determining these plausible alternatives varies between theories. For the Semantic Attraction account, the authors use a subjective criterion of semantic attractiveness based on experimenter intuition (they assume, for instance, that the alternative to *devouring* is *devoured*, and that this alternative will be attractive because a hearty meal can be devoured, but a dusty tabletop cannot be agent or patient for this verb). The Monitoring Theory defines plausibility of a sentence by computing semantic relatedness (LSA) between arguments and target verbs. CCA considers multiple subjectively-defined factors including animacy constraints, semantic association and sentence plausibility. In eADM, the notion of plausibility is loosely defined by a composition process based on "qualia" properties.

Others have proposed accounts involving a single processing stream that is responsible for integrating and using different sources of information. Prominent among single-stream accounts is that of Kuperberg (2016), which proposes that multiple hypothesized interpretations can be generated in parallel within a single processing system, and the interpretations compete to select a single interpretation based on reliability of many interacting cues.

Though multi-stream and single-stream models differ in assumptions on whether there are distinct processing mechanisms for syntactic and semantic cues, they share the basic intuition that a subset of cues may lead to a (possibly temporary) heuristic interpretation that can differ from the literal meaning. For all theories, this heuristic interpretation helps to account for the lack of N400 effect because the heuristic interpretation represents a more plausible interpretation, and the N400 response is hypothesized to reflect processing relative to this more plausible meaning. The P600, then, is typically considered to reflect a conflict resolution process between this plausible interpretation and the actual input—this can also be related to accounts of the P600 as reflecting perception of error, consistent with research linking the P600 to the domain-general P3b component triggered by less probable (non-linguistic) stimuli (Leckey & Federmeier, 2020). In the Semantic Attraction account, the heuristic interpretation is permanent, and the P600 reflects perceived grammatical error (which the plausible interpretation has corrected) in the input. For Monitoring Theory, the P600 reflects reprocessing to check for errors. For CCA, the P600 reflects a conflict between the semantic stream and one of the more literal streams. For eADM, the P600 reflects both conflict between different streams and a check for "well-formedness" of input.

While these theories provide compelling accounts, a central challenge that most of these theories face is the existence of biphasic N400-P600 effects (see e.g. Brouwer, Fitz, & Hoeks, 2012, for discussion). For instance, Chow et al. (2016) report a biphasic N400-P600 effect to *'The tenant inquired which exterminator the landlord had evicted.'*, relative to *'The tenant inquired which neighbor the landlord had evicted.'*. This effect is difficult to predict for the theories above because these theories tend to assume that for a given experimental manipulation, a more plausible heuristic interpretation either will or will not be available—and if it is available, the theories predict a P600 but no N400, whereas if it is not available, the theories predict an N400 but no P600. For some theories, like eADM, this challenge has motivated claims that the semantic P600 is generated by different mechanisms entirely.

Another challenge for these theories is that different ERP patterns can be observed in response to a seemingly identical linguistic manipulation. In particular, role reversal anomalies (*...which waitress the customer had served*) have famously been found to elicit a semantic P600

Table 1
A summary of psycholinguistic models for ERP effects.

Model	Main claim	N400 explanation	P600 explanation
Semantic Attraction	Semantic stream; syntactic stream	Plausibility in semantic stream	Detection of syntactic errors in the syntactic stream
Monitoring Theory	Heuristic stream; algorithmic stream	Plausibility in heuristic stream	Error monitoring
Continued Combined Analysis	Semantic stream; thematic stream; syntactic stream	Plausibility in semantic stream; can be blocked by conflict in other streams	Conflict resolution
Extended Argument Dependency Model	Semantic stream; thematic stream	Plausibility in semantic or thematic stream	Conflict resolution or well-formedness check
Kuperberg Generative Theory	Single stream with all cues	Plausibility of interpretation determined by strong cues at N400 time window	Updating previous interpretation

effect, with no immediate N400 effect (relative to canonical constructions, e.g., *...which customer the waitress had served*) (Chow, Lau, Wang, & Phillips, 2018; Chow et al., 2016; Yano, 2018). However, recent work testing with an identical role reversal manipulation has found an immediate N400 effect with no P600 effect (Ehrenhofer et al., in press), seemingly in direct contradiction of the earlier results. These conflicting results pose a challenge for the above theories because the theories typically predict existence or absence of a heuristic interpretation directly based on the experimental linguistic manipulation—for instance, predicting that a role reversal will have a readily available heuristic interpretation (the corresponding non-role-reversed interpretation), thus yielding no N400 effect—and therefore are not well-equipped to predict divergent results for an identical manipulation.

Our model is very close in spirit to these heuristic interpretation theories, in that we build directly on the intuition that the processor may entertain a more plausible heuristic interpretation that drives a lack of N400 effect. A key difference is that we formalize the process of arriving at the heuristic interpretation as a probabilistic selection between candidate interpretations, and this selection process – as well as quantification of N400 and P600 responses – is sensitive to statistical idiosyncrasies in the experimental stimuli, allowing for differing results depending on the particular stimuli. As a result, we find that our model is able to account for biphasic effects as well as divergent ERP patterns for a single linguistic manipulation, shedding light on particular mechanisms and sensitivities that can explain the observed patterns. Our account is largely agnostic as to whether these mechanisms should be considered to exist within a single stream or multiple streams—as we discuss further in Section 7, our formulation is in principle compatible with either of these frameworks, so we leave resolution of that dimension of inquiry for future work (see Table 1).

2.2. Computational models

A number of computational models have also been proposed to account for these ERP phenomena. Rabovsky, Hansen, and McClelland (2018) model the N400 as change in activations, within an internal neural network representation layer intended to capture the predicted meaning of the sentence. Brouwer, Crocker, Venhuizen, and Hoeks (2017), Brouwer, Delogu, Venhuizen, and Crocker (2021) propose a single-stream "Retrieval–Integration" model, in which the N400 reflects difficulty of lexical retrieval in context, and the P600 reflects difficulty of lexical integration in context. ERP responses are simulated as change of internal activation patterns in retrieval and integration layers of a neural network model (Brouwer et al., 2017), or as change of constructed utterance (surprisal) (Brouwer et al., 2021). Michalon and Baggio (2019) implement a multi-stream processing system where a semantic component uses lexical-semantic information to predict grammatical roles, while a separate syntactic component predicts the same set of grammatical roles with part-of-speech information. In this model, the N400 reflects detection of semantic error, operationalized as the accuracy of classifying whether the target verb can be a direct dependent of the subject in a corpus, and the P600 monitors mismatches between the syntactic labels predicted by the two components. While the Brouwer et al. (2017, 2021) models in particular differ from the above verbally-specified theories in having more of a capacity to handle biphasic N400-P600 patterns, all of these computational simulations use idealized synthetic inputs when training models—which means that like verbally-specified theories, they are not well-equipped to account for cases in which the same phenomenon produces different results. That is, these models operate in synthetic environments, the statistics of which are determined by planned variation introduced by experimenters—and this means that the training environments typically do not leave room for unplanned idiosyncrasies from other properties of stimuli.

In order to overcome limitations of synthetic stimuli, vector representations and neural networks trained on natural data have also been used for ERP modeling purposes. Ettinger (2018) uses word embeddings trained on natural corpora to account for stimulus idiosyncrasies, and is able to account for the divergent N400 responses to role reversals in Chow et al. (2016) and Ehrenhofer et al. (in press). However, this model simulates N400 amplitude with heavy reliance on cosine similarity between the target word and the last content word of the context—which means that it is poorly equipped to handle lexical changes that are more distant from the target in prior context. Consequently, this model will struggle with some of the additional experiments that we simulate in this paper. Other works have used neural networks trained on natural data for simulating the N400 component. Frank, Otten, Galli, and Vigliocco (2013, 2015) establish a basic correspondence between N400 amplitude and word surprisal estimated by recurrent neural networks (RNN). Michaelov and Bergen (2020) use RNN surprisal from real experimental stimuli to simulate N400 patterns from a range of neurolinguistic studies—finding, however, that this RNN surprisal is unable to predict N400 effects to morpho-syntactic and event structure violations. More recently, surprisal from transformers has been shown to outperform RNN surprisal in predicting N400 amplitude (Merkx & Frank, 2021; Michaelov, Bardolph, Coulson, & Bergen, 2021): these studies show that transformers account for more variance in N400 amplitude overall, but they do not test models' ability to simulate N400 patterns to specific target phenomena and linguistic manipulations as we do here. Other work (Lindborg & Rabovsky, 2021) has represented N400 amplitude as the difference between internal activation from the current word and the previous word in transformers (GPT-2), and has successfully simulated N400 effects in relation to semantic violations, cloze probability and unexpected primed words—however, this approach proves unsuccessful in simulating N400 blindness to role reversal anomalies. Fewer studies have attempted to simulate the P600 with neural network measures, and it has been observed that both surprisal and entropy reduction measures from RNNs may be inadequate to account for P600 or post-N400 positivities (Frank et al., 2015). Though neural network measures have proven useful for simulating aspects of ERP components

– specifically the N400 in certain settings – continuing limitations suggest that neural network measures alone may not be sufficient to explain the full range of ERP patterns. We propose to complement these neural network measures with a noisy-channel framework that is able to incorporate specialized mechanisms in distinct processing stages.

Noisy-channel models have been used for modeling syntactic ambiguity resolution (Levy, 2008), setting the precedent for modeling computations that may allow comprehenders to override the literal linguistic input. Gibson et al. (2013) show that a noisy-channel model can help to explain effects of stimulus properties on rates of non-literal interpretation in comprehension. Futrell, Gibson, and Levy (2020) also use noisy channel models to model language-dependent structural forgetting and locality effects. While these models incorporate noisy channel inference into various aspects of processing, they do not attempt to use this mechanism to account for ERPs. Most related to our work here is that of Ryskin et al. (2021), which proposes that amplitudes of the N400 and P600 components are indices of noisy-channel inference: as more errors in the stimuli are corrected, N400 amplitude is reduced and P600 amplitude increases. These authors design ERP experiments consisting of semantically and syntactically anomalous sentences, with an additional manipulation such that some sentences with semantic violations are more easily correctable (e.g., *The storyteller could turn any incident into an amusing* ***antidote/anecdote***), as measured by how likely humans are to correct the sentence into a control counterpart in an editing experiment. The authors find that N400 and P600 magnitudes are linearly related to ease of error correction, whether measured by word edit distance or by human accuracy in the error correction experiment. This model shares key intuitions with ours, in assuming an error correction process that can be linked to noisy-channel inference, and giving a probabilistic account that incorporates variation in error-correction behaviors at the stimulus level. However, we extend beyond what is done in that work, in implementing a full noisy-channel model to computationally quantify N400 and P600 amplitudes, taking into account finer-grained stimulus properties, and accounting for a broader range of psycholinguistic experiments. We provide additional detail about mechanistic differences between our model and that of Ryskin et al. (2021) in Section 5.

3. Simulated experiments

In the present paper we introduce a computational model and report the results of simulating N400 and P600 effects from nine psycholinguistic experiments with various types of semantic anomalies and syntactic anomalies. We select our experiments to cover the full range of ERP patterns that have been reported in response to the semantic violations of interest: N400-only, P600-only, and biphasic N400-P600 effects. Additionally, we simulate one experiment with a classic syntactic P600 effect, to show that effects to syntactic violations can be accounted for as well. In this section, we first review the experiments that we will be simulating.

3.1. Dataset

We use original stimuli from nine psycholinguistic experiments featuring semantic/thematic or syntactic violations, with empirical results varying between N400 effect only, P600 effect only, and biphasic N400-P600 effect (see Table 2). Our selected experiments include four kinds of linguistic manipulations: role reversal, animacy violation, word substitution and preposition deletion.

Role reversal. Two of our simulated experiments involve role reversal violations, in which the argument roles of nouns in sentences of the experimental condition are reversed relative to what would be expected in a canonical situation. In our simulations below, we include one experiment that reports a P600 effect and one that reports an N400 effect, to test the capacity of our model to explain both patterns of result. The first experiment, from Chow et al. (2016), we refer to as Reversal-1, and the second, from Ehrenhofer et al. (in press), we refer to as Reversal-2. In both of these experiments, the original stimuli are constructed using embedded object questions (e.g. *the restaurant owner forgot which customer the waitress had served...*) to ensure that information about argument roles precedes the verbs. ERP patterns are measured at the final verb position. To obtain role reversed sentences, the order of the two arguments within the embedded clauses is switched. The sentences with canonical thematic role assignment create highly probable scenarios (*. . . which customer the waitress had served...*), while sentences with reversed thematic roles violate comprehenders' world knowledge of thematic role assignment (*. . . which waitress the customer had served...*). For the Reversal-1 experiment, the authors report no N400 effect, but a significant P600 effect—consistent with previous findings on the semantic P600. By contrast, in the Reversal-2 experiment, which uses similar stimuli and experimental procedure, the authors report a larger N400 response in the role-reversed condition than in the control condition, but no significant P600 effect. As we describe above, this distinction is one of the primary sticking points for any account that allows for just one pattern of ERP effects given a role-reversal paradigm—this includes most verbally-specified theories of the semantic P600 reviewed above, as well as computational models that use idealized stimuli without access to idiosyncrasies of real experimental items.

Animacy. The second set of simulated studies consists of three experiments involving violations of animacy constraints on noun phrases. The experiments that we refer to as Animacy-1 and Animacy-2 are drawn from Kim and Osterhout (2005) and the experiment that we refer to as Animacy-3 is drawn from Kuperberg et al. (2010). For all of these experiments, the subject of the sentence is inanimate, while the target verb requires either an animate subject (*critical condition*) or an animate object (*control condition*). Whether it is the subject or object that must be animate is manipulated by use of the active (e.g. *the hearty meal was devouring...*) or passive voice (e.g. *the hearty meal was devoured...*) with verbs that require animate agents. The critical distinction between Animacy-1 and Animacy-2 is the degree of lexico-semantic association, which refers to the degree to which the target verb is semantically/thematically associated with the arguments in the context. Though semantic association is thought not to be an essential trigger for the semantic P600, experiments in this set have suggested that the semantic P600 effect is more likely when there is a strong semantic association between the verbs and the arguments. In Animacy-1, subjects are semantically attractive as patients/themes for the target verbs, such that these sentences can easily be revised to more likely alternatives by changing the grammatical form of the verb (e.g., *The hearty meal was devouring → The hearty meal was devoured*). This experiment shows a classic semantic P600: no N400 effect, but a P600 effect. In Animacy-2, there is significantly lower lexico-semantic association between the subject and the target verb (e.g., *The dusty tabletops were devouring/devoured*). In this experiment, the authors report an N400 effect, but no P600 effect. In Animacy-3, the lexico-semantic association between verbs and subjects varies without being explicitly manipulated (the Animacy-3 experimental and control conditions were intended to be compared against a third, complement coercion condition, which is not included in our experiments).[1] In the Animacy-3 experiment, the authors report a biphasic N400-P600 effect on critical verbs.

[1] There are three conditions in the original Animacy-3 experiment: *the journalist began/wrote/astonished the article*. The ERP effects in the coercion condition (*began*) and in the animacy violation condition (*astonished*) are both compared with the control condition (*wrote*). The study shows that the N400 effect in the coercion condition is similar to the N400 in the animacy violation condition, suggesting that the verb's semantic structure is stored at a different level from syntactic structure.

Table 2
List of simulated experiments, with experimental manipulations and results. The underlined word is used in the canonical condition. The italicized word in parentheses is used in the critical (experiment) condition. The target word is marked as bold. Presence of N400/P600 effect is determined by difference of ERP amplitudes between canonical and critical conditions.

Experiment	Manipulation	Sample sentence	Observed effect
Reversal-1	Role-reversal	The restaurant owner forgot which customer (*waitress*) the waitress (*customer*) had **served**...	P600
Reversal-2	Role-reversal	...which bull (*cowboy*) the cowboy (*bull*) had **ridden** out on the range.	N400
Animacy-1	Active/passive	The hearty meal was **devoured** (***devouring***)...	P600
Animacy-2	Active/passive	The hearty meal (*dusty tabletops*) were devoured (*devouring*)...	N400
Animacy-3	Active/passive	The journalist wrote (*astonished*) the **article**.	Biphasic
Substitution-1	Word substitution	The tenant inquired which neighbor (*exterminator*) the landlord had **evicted**...	Biphasic
Substitution-2	Word substitution	The tenant inquired which neighbor (*exterminator*) the landlord had **evicted**...	Biphasic
Substitution-3	Word substitution	The exterminator (*neighbor*) inquired which neighbor (*exterminator*) the landlord had **evicted**...	N400
Preposition-1	Preposition deletion	Kim recommended Shakespeare to (∅) **everyone**...	P600

Word substitution. The third set of experiments that we simulate involve word substitution. In these experiments, an argument noun within an embedded clause is either changed to a different word that fits less well in the context, or swapped with the main subject noun. In Substitution-1 and Substitution-2, both from Chow et al. (2016), the authors substitute one of the arguments in the embedded clause (e.g., *...which exterminator the landlord had evicted...* → *...which neighbor the landlord had evicted...*), which significantly lowers the semantic association between the target verb and the preceding context. Substitution-2 shares a portion of the stimuli from Substitution-1, but the stimuli in Substitution-2 are expanded by pairing both contexts with a probable continuation (e.g. *which neighbor the landlord had evicted; which exterminator the landlord had hired*). For both Substitution-1 and Substitution-2, the authors report a biphasic N400-P600 effect. In Substitution-3, also from Chow et al. (2016), the authors switch one argument in the embedded clause with the main subject (e.g. *The neighbor inquired which exterminator the landlord had evicted...* → *The exterminator inquired which neighbor the landlord had evicted...*), thus keeping lexical content the same between conditions. In the Substitution-3 experiment, the authors report only an N400 effect.

Preposition. The last experiment that we simulate involves manipulation of syntactic violations by deleting the syntactic preposition (*to*) in a dative construction with an indirect object (*Kim recommended Shakespeare to everyone...*) (Ainsworth-Darnell et al., 1998). The authors of this experiment report a P600 effect with no N400 effect—we include this experiment as a representative of the classically observed syntactic P600 effect.

3.2. Use of real stimuli

Compared with many computational psycholinguistic models that have attempted to account for these types of phenomena while using idealized synthetic inputs, a key distinction of our model comes from the use of real experimental stimuli, which allows us to account for idiosyncratic properties of individual items. While psycholinguistic experiments typically manipulate a small set of key linguistic properties and link the manipulated features to one target cognitive question, it is difficult to control all variables that may influence the outcomes. The characteristics of stimuli from experiments with highly similar linguistic structures and the same experimental design may still differ in critical variables, which can give rise to different ERP results, as seen in the divergent outcomes for Reversal-1 and Reversal-2. It has already been observed that the preceding contexts and predicted verbs have closer lexical associations in Reversal-1 than in Reversal-2, as measured by cosine similarities obtained from GloVe embeddings (Ettinger, 2018). We argue that psycholinguistic models can be strengthened by taking such item-level properties into consideration—otherwise models will only be able to offer one prediction for a group of experiments with the same linguistic manipulation (see Table 3). As we will demonstrate below, the use of real experimental stimuli can also offer explanations for biphasic effects that are not apparent from the nature of the experimental manipulations.

4. Model overview: a noisy-channel based model of ERPs to semantic anomalies

In this section we provide a high-level overview of the model that we propose to account for the observed ERP effects (see Fig. 1). In this model, a noisy channel computation decides the heuristic interpretation for a given sentence input, balancing interpretation plausibility with extent of revision. The N400 then reflects the fit of the target word within the interpretation computed through that noisy channel model, while the P600 reflects level of conflict or divergence between the heuristic and literal interpretations. We incorporate quantitative estimates for measures like plausibility and semantic similarity for each of the individual experimental stimuli, by using proxy measures from pre-trained neural network models. Our model builds closely on existing intuitions that processing of these sentences should involve a mechanism for forming heuristic interpretations driven by plausibility—however, our model formalizes this process as a probabilistic candidate selection within a noisy-channel framework, and leverages large pre-trained neural network models to enable sensitivity to idiosyncratic properties of individual stimuli.

4.1. Computing heuristic interpretation: the noisy channel model

Our model assumes a stage of comprehension involving computation of an initial heuristic interpretation—this heuristic interpretation weighs the prior plausibility of each potential interpretation against the likelihood of perceiving the input sentence given the potential interpretation. Note that we do not make any specific assumption about whether this heuristic interpretation persists as an error correction through the final comprehension stage, or serves only as a temporary early interpretation to be replaced later by a more literal interpretation. For the purpose of our model, it is only necessary that both heuristic and literal interpretations exist, since divergence between heuristic and literal interpretations will drive our simulated P600.

We formulate the heuristic interpretation process as a selection process over candidate interpretations, using a noisy channel model. The noisy channel computation assumes that the speaker chooses an intended sentence m to convey their intended meaning, but that the perceived sentence s reflects potential distortion due to noise. The noisy

Table 3
Predicted ERP patterns for various psycholinguistic models.

Model	Reversal-1,2	Animacy-1,3	Animacy-2	Substitution-1,2,3	Preposition-1
Semantic Attraction	no effect	P600	N400	N400	P600
Monitoring Theory	P600	P600	N400	N400	P600
CCA	P600	P600	N400	N400	P600
eADM	P600	P600	N400	N400	P600
Generative Theory	P600	P600	N400	N400, (P600)[a]	P600

[a] The presence of the ERP effect depends on the specific cues and relative strengths in the model setup.

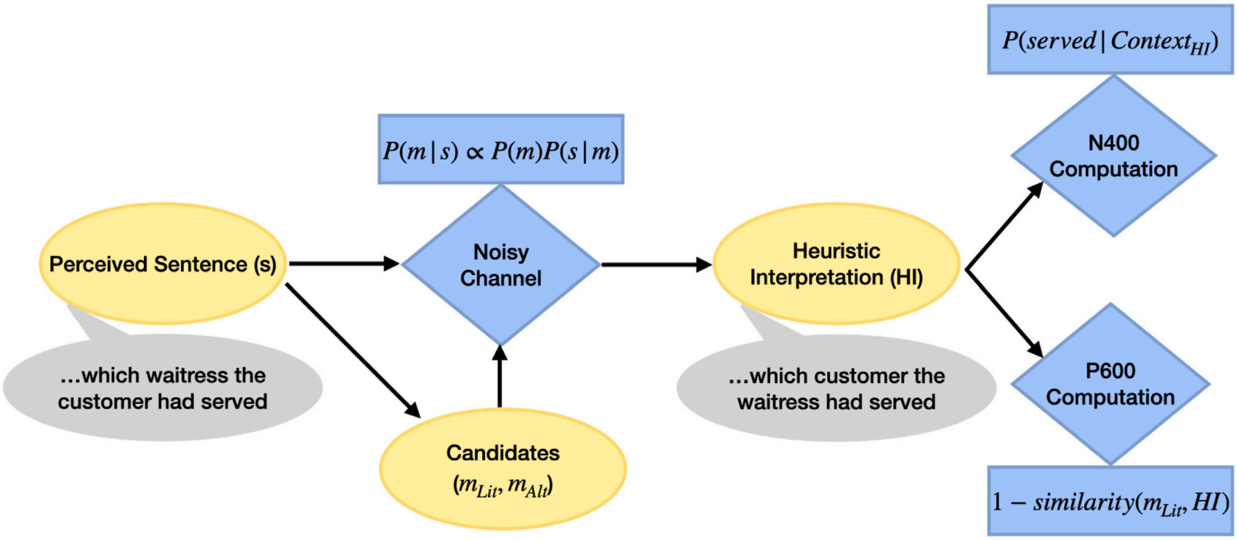

$P(served \mid Context_{HI})$
$P(m \mid s) \propto P(m)P(s \mid m)$
N400 Computation
Perceived Sentence (s)
Noisy Channel
Heuristic Interpretation (HI)
...which waitress the customer had served
...which customer the waitress had served
P600 Computation
Candidates (m_{Lit}, m_{Alt})
$1 - similarity(m_{Lit}, HI)$

Fig. 1. A overview of model architecture. Candidates (m) are operationalized as literal interpretation (m_{Lit}) and alternative interpretation (m_{Alt}).

channel computation involves Bayesian inference of the true intended sentence (m) given the perceived stimulus (s).

$$P(m \mid s) = \frac{P(m)P(s \mid m)}{P(s)} \tag{1}$$

In this computation, the prior $P(m)$ represents the prior probability of the interpretation under consideration, independent of its relationship to the perceived sentence. As we describe in greater detail below, we aim for the computation of this prior probability to reflect a general notion of plausibility of the interpretation m, using probabilities from a pre-trained neural network language model.

The likelihood $P(s|m)$ in this computation represents how likely it is that the interpretation m would be distorted into the perceived sentence s during message transmission. This ensures that interpretations that are extremely different from the perceived sentence will not receive high probability. As we describe below, for computing this value we use a simple edit distance measure.

We consider the outcome of the heuristic interpretation process to be selection of a single interpretation HI, which corresponds to the interpretation m with the greatest posterior probability $P(m|s)$, selected from a set M_s of possible interpretations for a given stimulus s, as in (2):

$$\begin{aligned} \text{HI} &= \operatorname*{argmax}_{m \in M_s} P(m \mid s) = \operatorname*{argmax}_{m \in M_s} \frac{P(m)P(s \mid m)}{P(s)} \\ &= \operatorname*{argmax}_{m \in M_s} P(m)P(s \mid m) \end{aligned} \tag{2}$$

If the chosen heuristic interpretation differs from the literal meaning m_{Lit} of the presented sentence, we consider there to have been a (possibly temporary) “error correction” during the heuristic interpretation process. Alternatively, if the interpretation with the highest posterior matches the literal meaning of the presented sentence, we consider there to have been no error correction.

4.2. Computing ERP components

Within our model, the simulation of both N400 and P600 amplitudes depends upon the nature of the heuristic interpretation.

N400. In line with many classic psycholinguistic theories, in our model the N400 can be considered to reflect sensitivity to semantic anomaly. Specifically, we compute N400 amplitude based on fit of the target word within the heuristic interpretation. This means that correction of an anomalous presented sentence to a more plausible heuristic interpretation will typically result in reduced N400 amplitude in the anomalous condition (due to better fit of the target word), which can result in lack of N400 effect between anomalous and control sentences. This is consistent with many of the multi- and single-stream models described above, where the existence of a plausibility-driven interpretation is assumed to account for a lack of N400 effect. As we describe below, we compute fit of the target word using conditional probability from a pre-trained neural network language model.

P600. Also consistent with many psycholinguistic theories, our model considers the P600 to reflect effort to reconcile conflict between different interpretations. We assume that the P600 reflects access to an interpretation corresponding to the literal meaning of the presented sentence, in addition to the heuristic interpretation which may deviate from the literal meaning. P600 amplitude is then computed based on the amount of semantic divergence between the heuristic interpretation and the literal interpretation, to reflect effort of reconciliation between these interpretations. If comprehenders derive a heuristic interpretation that is different from the literal meaning, this will require more substantial reconciliation between these interpretations, leading to a larger P600. As we describe below, we compute semantic divergence using a pre-trained neural network model trained on semantic similarity. (Note again that our model does not assume a specific temporal relationship between the heuristic and literal interpretations—only that the processor has access to both interpretations at the point at which the P600 is generated, and to the heuristic interpretation at the point at which the N400 is generated.)

5. Model implementation

Here we describe the details of our implementation of the model summarized above.

5.1. Noisy channel model (selection of heuristic interpretation)

As described above, we compute posterior probabilities for candidate heuristic interpretations m based on (1) the prior probability of m, and (2) the likelihood of seeing s as a distortion of m.

Prior. For the prior probability $P(m)$, we aim to capture a version of interpretation plausibility, which we approximate via sentence probability estimates from a large neural network pre-trained on word prediction (OpenAI GPT) (Radford, Narasimhan, Salimans, & Sutskever, 2018).[2] This model is a language model, which means that it assigns probabilities to sequences of words, and to words in context. The measure that we use for our prior is inverse perplexity of a sentence based on probabilities from the neural network language model. Perplexity is a standard measure for evaluation of language models, which can be thought of as quantifying how surprising the sequence is based on the model's probabilities. If the model assigns generally high probabilities to the words within the sequence, the perplexity will be lower. For our prior we use inverse perplexity, so that sequences that the model finds less surprising (e.g., more plausible, or more grammatical sentences) will receive higher priors.

$$P(m) \propto \frac{1}{\text{Perplexity}(m)} = P(w_1, w_2, \ldots, w_n)^{\frac{1}{n}} \tag{3}$$

To what extent do the perplexities from this neural network actually embody a notion of plausibility? The flexibility of these neural network models enables us to derive fine-grained measures for real experimental stimuli, but the trade-off is some reduction in transparency. To check the relationship of our model's prior measure with standard notions of plausibility, we compare average prior values assigned by the model to items from experimental and control conditions from the experiments to be simulated in this paper. Given that in the majority of these simulated experiments the items in the experimental condition are designed to contain semantic anomalies, we can expect that ground truth plausibility in the experimental condition should be lower than in the experimental condition. Fig. 2 shows the results. We see that the model consistently assigns lower prior probability (higher perplexity) to experimental conditions relative to control conditions, despite the fact that sentence pairs differ only slightly in surface form between conditions. We also see significant variation in the magnitude of the prior difference between conditions; as we will discuss in the analysis of results below, this variation will be relevant in accounting for the range of ERP results. The patterns observed here support the ability of this neural network measure to capture the basic plausibility differences expected between experimental and control conditions, and also previews how variation in this measure can help in capturing differences between experiments. We can also compare these priors against offline human plausibility ratings in Table 5, where we see several points of alignment with model values: (1) greater plausibility difference in Reversal-1 than Reversal-2, (2) experimental sentences in Animacy experiments show the lowest plausibility and the greatest difference from control conditions, and (3) the experimental condition in Substitution-2 shows higher plausibility than in the other two Substitution experiments, whereas the control sentences in Substitution-1 are rated more plausible than in the other experiments. Overall, these patterns give us good reason to believe that our priors represent solid proxies for effects of plausibility. Note, however, that while we see correspondence between our estimate of plausibility and human offline plausibility judgments, our model's notion of plausibility is not identical to that embodied by human offline ratings. We discuss this divergence further in Section 7, but one immediately noteworthy deviation from standard notions of plausibility is the fact that our measure is sensitive not only to semantic anomalies, but also to the syntactic anomaly in Preposition-1. This fact will be important in successfully simulating that experiment, and it raises the important possibility that the relevant "plausibility" mechanisms underlying these effects are in fact broader than intuitive notions of plausibility as embodied in offline judgments.

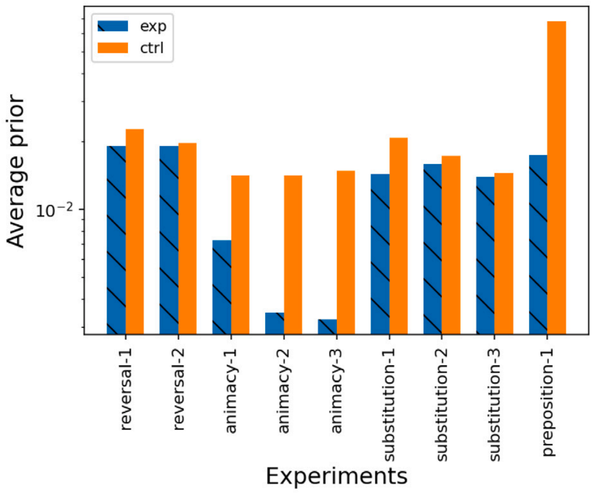


Fig. 2. Average prior (log-scale) of literal sentence interpretation in experimental and control conditions across experiments.

Likelihood. We base the likelihood $P(s|m)$ on the Damerau–Levenshtein edit distance (D) between m and s, to capture the greater likelihood of interpretations that represent smaller deviations from the true input.[3] We design our edit distance metric such that it is more costly to do insertion and substitution (cost 2) than deletion and swap (cost 1), in line with experimental evidence showing that comprehenders are more likely to correct sentences if the correction requires deletion or swap than if it requires insertion (Gibson et al., 2013; Poppels & Levy, 2016; Ryskin, Futrell, Kiran, & Gibson, 2018). We calculate string distance similarity based on the number of characters in the longest string (Max), and the weighted edit distance (D). The likelihood measure is then defined as follows:

$$P(s \mid m) \propto 1 - \frac{D(m, s)}{Max(m, s)} \tag{4}$$

Fig. 3 shows average likelihood $P(s|m)$ in the experimental condition, for candidate interpretations that differ from the literal interpretation m_{Lit}. (For the literal interpretation m_{Lit}, the likelihood is always 1.) We see that the likelihood for Preposition-1 is highest, as it only requires deletion of two characters (*to*). Animacy-1 and Animacy-2 also show higher likelihoods than most experiments, because the construction of alternative sentences in these two cases only

[2] We also tried priors estimated from GPT-2 (Radford, Wu, Child, Luan, Amodei, Sutskever, et al., 2019) and BERT (Devlin, Chang, Lee, & Toutanova, 2018), among which GPT provides sentence probabilities most similar to human plausibility ratings.

[3] We also tried unweighted Damerau–Levenshtein edit distance, and weighted and unweighted Levenshtein edit distance—we find that these variations do not make any major difference in the final results.

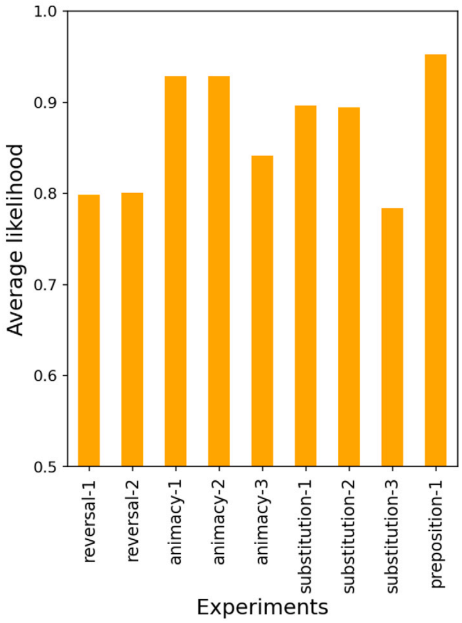


Fig. 3. Average likelihood (string distance similarity) between literal and alternative interpretations in critical conditions across experiments.

involves modification of morphemes (e.g. *-ed* → *-ing*), whereas the other experiments require either change of position for two words or substitution/insertion of a full word. Across experiments, likelihood tends to be similar, as our alternative interpretations are chosen to have minimal changes from the literal sentence—as a result, likelihood is not the primary driving force in variation of ERP patterns across experiments in our simulations. For all experiments, the likelihood encourages the interpretation to stay consistent with the presented sentence and penalizes letter-wise modifications, which discourages error correction unless the gain in plausibility is sufficient.

Alternatives. To simplify model computation, for each stimulus we limit to two candidate intended messages, of which one is the literal interpretation m_{Lit} (*Lit* in Table 4). The other candidate is an alternative interpretation m_{Alt} (*Alt* in Table 4) derived manually following principles often inspired by the original experimental manipulations, but with the primary goal of identifying a minimal change that results in a plausible alternative for the anomalous item. The same alternation type that we select for the anomalous items is then also applied for the control items (such that the control items typically have a slightly more anomalous interpretation as the non-literal candidate). Table 4 shows examples of these alternations. To form the alternative interpretation for Reversal experiments, the position of the two arguments in the embedded clause is simply swapped. For Animacy experiments, the form of the target verb is changed from past participle to progressive or vice versa, switching the sentence between passive and active voices. For Substitution experiments, one argument is replaced with another lexical item that appears in the same position of the counterpart experimental stimulus. For Preposition-1, the syntactic preposition *to* is inserted or deleted. In most cases, the alternatives align with manipulations from the original psycholinguistic experiment, such that the alternative for the experimental condition is the corresponding sentence in the control condition, and vice versa. The exceptions are Animacy-2 and Animacy-3, in which we alternate the sentences between active and passive voice, instead of the word substitution manipulation done in the original experiments. We make this exception for better consistency with construction of alternatives in the other animacy-based experiment, and because this is a smaller change to derive a plausible alternative interpretation.

As described above, we use the noisy channel model to compute posterior interpretation probabilities for each candidate interpretation of a given stimulus, and consider the candidate with the highest posterior probability to be the selected heuristic interpretation HI.

5.2. Simulating N400 amplitude

Our N400 computation is designed to capture difficulty of processing a target word in context, assuming that this processing reflects the selected heuristic interpretation. Specifically, we compute N400 amplitude based on the conditional probability of the target word in the context of the heuristic interpretation. These conditional probabilities are again derived from GPT (Radford et al., 2018). When the target word fits well with the heuristic context, the conditional probability $P(\text{target} \mid \text{context}_{\text{HI}})$ is greater—resulting in smaller N400 amplitude, which is computed as below:

$$\text{N400 amplitude} = -P(\text{target} \mid \text{context}_{\text{HI}}) \tag{5}$$

Under this model, the N400 amplitude to target words will be smaller if the targets have better fit to the selected heuristic context. The presence of an N400 effect is determined by the difference of average N400 amplitude between experimental and control conditions.

How reasonable are the conditional probabilities produced by the neural network model? Recall that these probabilities come from the same model as the prior (plausibility) measure above—however, by contrast to that perplexity measure, this conditional probability measure focuses on how surprising the target word is in context, rather than the probability of the full sentence. Fig. 4 shows the conditional probabilities for sentences from the experimental and control conditions across experiments. Again, we see that the conditional probability in the control condition is consistently greater than in the experimental condition, indicating that based on these estimates, the target in the control condition has better fit to context than in the experimental condition, as we would expect.[4] The exception to this pattern is Preposition-1, where we see that the syntactic anomaly has negligible effect on target word fit to context, by contrast to the large effect of syntactic anomaly on probability of the full sentence (Fig. 2).

We note that these conditional probabilities obtained from the neural network do deviate from patterns seen in cloze probabilities (derived from human fill-in-the-blank responses). For instance, in Fig. 4 we see that for Reversal experiments, the model assigns comparably high conditional probabilities in both experimental and control conditions. By contrast, Table 5 shows that cloze probabilities calculated from human responses have the targets in experimental conditions at cloze probabilities of approximately zero. Cloze probabilities are not available for Preposition-1, but we might expect a larger cloze probability difference between conditions here as well, given that the target in the experimental condition is ungrammatical. It has been previously observed that conditional probabilities estimated from neural networks tend to overestimate the probability of anomalous targets that share semantic association with the context, or that are syntactically related to the probable continuation (LeBrun, Sordoni, & O'Donnell, 2021; Michaelov & Bergen, 2020), relative to cloze probabilities. As a result, the conditional probability will predict a smaller N400 effect to "attractive" anomalies as compared to cloze probability. We will discuss more about the role of this divergence in Section 7.

[4] We see in Fig. 4 that certain conditional probabilities are especially low—for instance, Animacy-1/Animacy-2 experimental conditions. We note that these two conditions all use the gerund form for the target word, and inspection shows that these conditions also contain a high percentage of items in which the target is preceded by "had been". This structure seems to drive the particularly low conditional probabilities in those conditions.

Table 4
Sample candidate interpretations for items of all simulated experiments. Words in italics are changed in order to derive alternative (non-literal) interpretations as minimal changes from the literal interpretations. The target word is marked in bold.

Experiment	Critical condition	Control condition
Reversal-1	*Lit*: ... which *waitress* the *customer* had **served**... *Alt*: ... which *customer* the *waitress* had **served**...	*Lit*: ... which *customer* the *waitress* had **served**... *Alt*: ... which *waitress* the *customer* had **served**...
Reversal-2	*Lit*: ...which *cowboy* the *bull* had **ridden**... *Alt*: ...which *bull* the *cowboy* had **ridden**...	*Lit*: ...which *bull* the *cowboy* had **ridden**... *Alt*: ...which *cowboy* the *bull* had **ridden**...
Animacy-1	*Lit*: The hearty meal was ***devouring*** ... *Alt*: The hearty meal was ***devoured***...	*Lit*: The hearty meal was ***devoured*** ... *Alt*: The hearty meal was ***devouring***....
Animacy-2	*Lit*: The dusty tabletops were ***devouring***.. *Alt*: The dusty tabletops were ***devoured***..	*Lit*: The hearty meal was ***devoured***... *Alt*: The hearty meal was ***devouring***...
Animacy-3	*Lit*: The journalist *astonished* the **article**. *Alt*: The journalist *was astonished by* the **article**.	*Lit*: The journalist *wrote* the **article**. *Alt*: The journalist *was written by* the **article**.
Substitution-1	*Lit*: The tenant inquired which *exterminator* the landlord had **evicted**... *Alt*: The tenant inquired which *neighbor* the landlord had **evicted**...	*Lit*: The tenant inquired which *neighbor* the landlord had **evicted**... *Alt*: The tenant inquired which *exterminator* the landlord had **evicted**...
Substitution-2	*Lit*: The tenant inquired which *exterminator* the landlord had **evicted**... *Alt*: The tenant inquired which *neighbor* the landlord had **evicted**...	*Lit*: The tenant inquired which *neighbor* the landlord had **evicted**... *Alt*: The tenant inquired which *exterminator* the landlord had **evicted**...
Substitution-3	*Lit*: The *neighbor* inquired which *exterminator* the landlord had **evicted**... *Alt*: The *exterminator* inquired which *neighbor* the landlord had evicted.	*Lit*: The *exterminator* inquired which *neighbor* the landlord had **evicted**... *Alt*: The *neighbor* inquired which *exterminator* the landlord had evicted.
Preposition-1	*Lit*: Kim recommended Shakespeare **everyone**... *Alt*: Kim recommended Shakespeare *to* **everyone**...	*Lit*: Kim recommended Shakespeare *to* **everyone**... *Alt*: Kim recommended Shakespeare **everyone**...

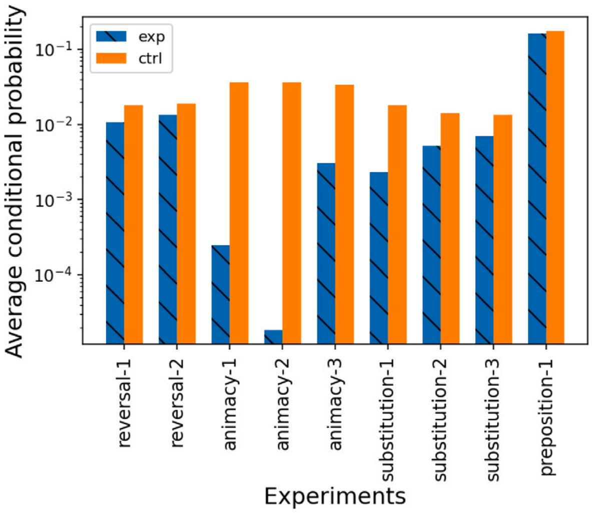


Fig. 4. Average conditional probability of target (log-scale) in literal interpretation for experimental and control sentences across experiments.

Table 5
Human plausibility ratings and cloze probabilities for simulated experiments, as reported in original human experiments. Plausibility ratings are converted into a scale of 0–100.

	Plausibility rating		Cloze probability	
	experiment	*control*	*experiment*	*control*
Reversal-1	23.8	85.4	0	.25
Reversal-2	40	87.9	<.02	.36
Animacy-1	3	92	NA	NA
Animacy-2	6	92	NA	NA
Animacy-3	NA	76	NA	.14
Substitution-1	31.1	90.3	0	0.28
Substitution-2	46.7	83.1	.004	.22
Substitution-3	29.2	77.2	.008	.22
Preposition-1	NA	NA	NA	NA

5.3. P600 simulation

To capture reconciliation between interpretations, we simulate P600 amplitude as the amount of difference between the selected heuristic interpretation HI and the literal interpretation m_{Lit}. We quantify this difference by extracting vector representations of the sentences (heuristic and literal interpretations) from a neural network trained to detect semantic similarity (fine-tuned DistilBERT) (Reimers & Gurevych, 2019), and computing the cosine similarity between these vector representations. Below is the computation of the P600 amplitude, where V denotes the function mapping a sentence to its vector representation.

$$\mathrm{P600\ amplitude} = 1 - \mathrm{cosine}(\mathrm{V(HI)}, \mathrm{V(m_{Lit})}) \tag{6}$$

Under this model, we can say roughly that P600 amplitude will be greater if the heuristic interpretation has greater semantic divergence from the literal interpretation, and will be smaller if the semantic representation of the heuristic interpretation is closer to that of the literal interpretation. The presence of a P600 effect is determined by the difference in average P600 amplitude between experimental and control conditions.

For checking the validity of our semantic divergence measure, we have less of a clear a priori comparison than we have with the plausibility and N400 measures. However, we can spot check the similarities assigned by the measure to different sentence pairs. Fig. 5 shows the average cosine similarities for each experiment. For each experiment these cosines are computed between literal and alternative interpretations of the sentences in the experimental condition, so these similarities will be reflected in the P600 in the experimental condition (for a given stimulus) if the noisy channel model opts to correct the input sentence to the more plausible counterpart. We see that sentences differing only in relative position of content nouns (Reversal-1 and -2, Animacy-3, Substitution-3, Preposition-1) have very high similarity. Sentences differing in inflection of the verb (Animacy-1, Animacy-2) also have generally high similarity. Completely changing a noun of the sentence (Substitution-1, -2) yields a lower cosine similarity, suggesting that the neural network measure places particular weight on word content. Fig. 6 shows cosine similarity between literal and alternative interpretations for all items in experimental conditions. We see that there is greater variation for experiments with word substitution (Substitution-1, -2). Animacy-2 also shows some variation, which we speculate may depend on how much the dominant reading of the verb

Table 6
Sample cosine similarities between critical and alternative sentences.

Experiment	Experimental (*Alternative*) sentences	Similarity
Reversal-1	... which journalist (*celebrities*) the celebrities (*journalist*) had interviewed.	0.97
Animacy-1	The legal contract had been signing (*signed*).	0.96
Animacy-2	The man's signature was forging (*forged*).	0.73
Animacy-3	The composer astonished (*wrote*) the song.	0.97
Substitution-1,2	The tenant inquired which exterminator (*neighbor* the landlord had evicted.	0.88
Substitution-3	The neighbor (*exterminator*) inquired which exterminator (*neighbor* the landlord had evicted.	0.95
Preposition-1	Kim recommended Shakespeare ∅ (*to*) everyone.	0.97

Table 7
Sample cosine similarities between random sentence pairs.

Sentence 1	Sentence 2	Similarity
The librarian documented which journalist the celebrities had interviewed.	The composer astonished the song.	0.15
The legal contract had been signing.	The librarian documented which celebrities the journalist had interviewed.	0.19
The mans signature was forging.	The exterminator inquired which neighbor the landlord had evicted.	0.12

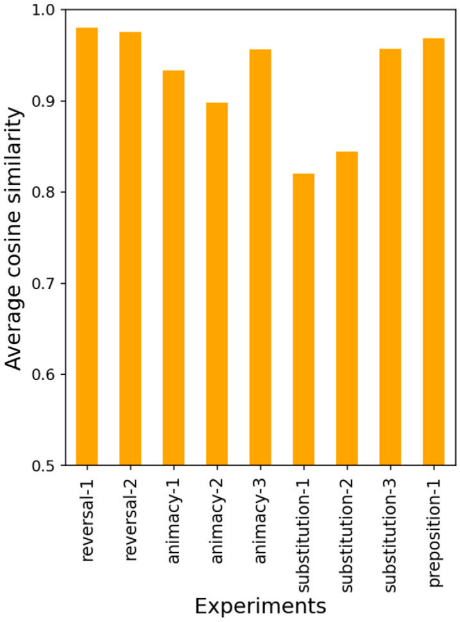
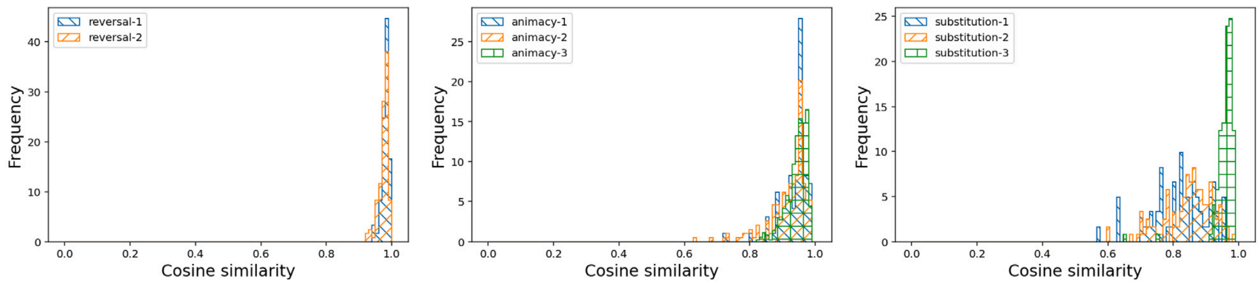


Fig. 5. Average cosine similarity (log-scale) between literal and alternative sentences in experimental conditions across experiments (strength of P600 amplitude when error correction occurs).

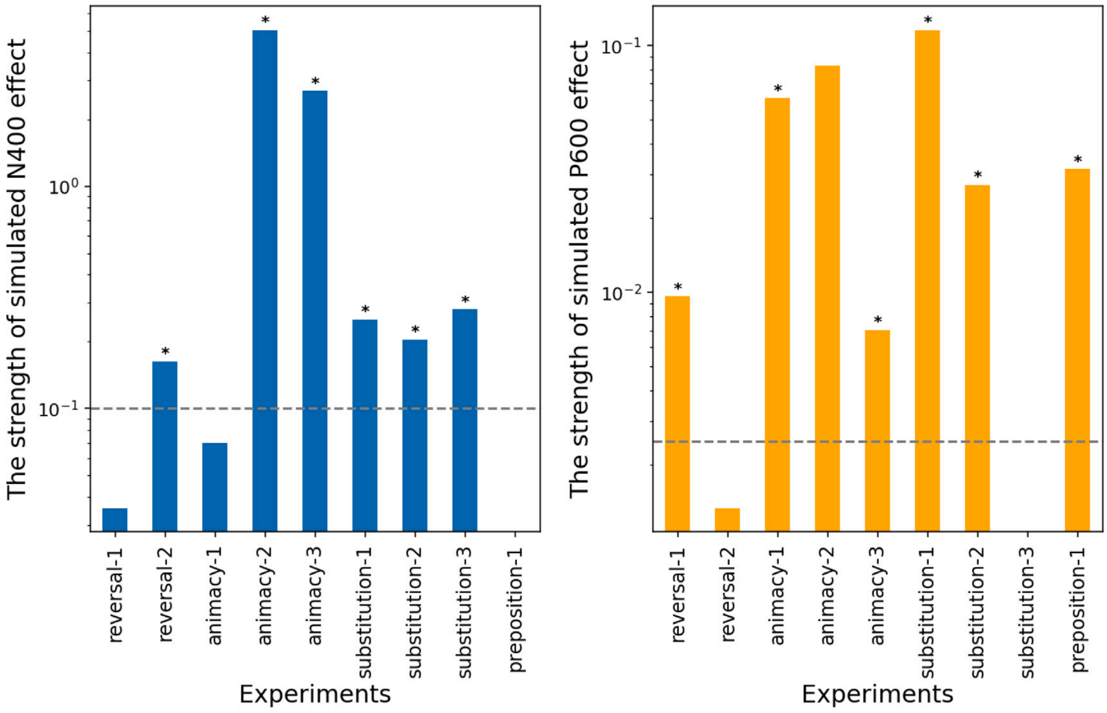
changes with the change in inflection. Table 6 shows cosine similarities assigned by the model to specific pairs from each of the simulated experiments. Note the lower cosine similarity on the pair from Animacy-2, which may be attributable to verb polysemy (e.g., *forging* may be most commonly used in contexts like *forging ahead* while *forged* may more commonly be used to refer to forged currency). In general, note that cosine similarities between our literal and alternative sentences are quite high compared to non-minimal sentence pairs (samples of which are shown in Table 7).

It is worth highlighting at this point the details of the distinctions between our simulation of the N400 and P600 and that in Ryskin et al. (2021). In Ryskin et al. (2021), the amplitudes of both the N400 and P600 are linked directly to the probability of error correction. Specifically, both components are predicted based on the ratio of the posterior probability of an alternative interpretation relative to a literal interpretation, given a presented sentence ($\frac{P(m_{Alt}|s)}{P(m_{Lit}|s)} = \frac{P(m_{Alt})P(s|m_{Alt})}{P(m_{Lit})P(s|m_{Alt})}$). When this posterior probability ratio becomes higher, the N400 is predicted to be smaller, and the P600 is predicted to be larger. In our model, the N400 is indexed by target word conditional probability given the heuristic interpretation, while the P600 is indexed by semantic divergence between heuristic and literal interpretations, where the heuristic interpretation is selected between a literal and a more plausible alternative interpretation via a noisy-channel process. So although these two modeling approaches share the intuition that the N400 and the P600 are related to error correction of a presented sentence into a more plausible alternative, they differ in the role and influence of this error correction. In Ryskin et al. (2021), ERP amplitudes directly reflect the size of the posterior probability ratios between interpretations, while our ERP amplitudes do not—in our model, as long as the posterior ratio is sufficient to trigger error correction, the size of that ratio will have no further role. Our N400 will instead reflect the fit of the target word given the selected heuristic interpretation, and our P600 will reflect semantic divergence between the heuristic and literal interpretations. Though we can expect some level of correlation between the posterior probability ratios and our measures of conditional probability and semantic divergence, these two sets of measures will not produce equivalent patterns—and additionally, these computational distinctions have corresponding differences in the theoretical cognitive implications of our models. In particular, our use of target word conditional probability and semantic divergence hypothesize that the N400 and P600 components are not driven by a noisy channel process alone, but rather by separate cognitive mechanisms that act on the outcome of a noisy channel process. Specifically, our model suggests that the N400 is driven by mechanisms influenced by target word fit to heuristic context, like pre-activation by, or integration of target word information into, the heuristic interpretation—while the P600 reflects a mechanism of resolving conflict between heuristic and literal interpretations, likely linked to updating mental representations to form a coherent situation model of sentence meaning. The success that our model shows in simulating a wide range of N400 and P600 results lends some credence to the mechanisms posited in our account—however, it is also possible that additionally incorporating a more direct role for posterior probability ratios could further enhance the model's alignment with human patterns. This will be a useful possibility to investigate in future work. As a final note, given that our model's



(a) Reversal (b) Animacy (c) Substitution

Fig. 6. Frequency distribution of cosine similarity between each critical and alternative sentence pairs in experimental condition across experiments.



(a) N400 (b) P600

Fig. 7. Simulated N400 (left) and P600 (right) effects (log-scale) 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.

simulation of the P600 quantifies reconciliation in terms of "semantic divergence", it is natural to wonder whether we will potentially underestimate certain syntactic P600 effects. However, as we show below, our model does predict P600 effects to classic syntactic anomalies. This suggests that our divergence measure may incorporate elements that fall outside of traditional notions of semantic divergence—we provide further discussion of this measure and its implications and limitations in Section 7.

6. Simulation results

We now present the results of applying this model to the experimental stimuli from our nine selected experiments. The model simulation results are shown in Fig. 7. In the figure we have plotted the difference between the average amplitudes in the control and experimental conditions for the experiment in question (see Eq. (7)), as in Eq. (7).

$$\text{Simulated ERP effect} = \overline{\text{ERP}}_{\text{exp}} - \overline{\text{ERP}}_{\text{ctrl}} \tag{7}$$

Asterisks in the figure indicate experiments in which there was a significant effect in the original human experiment, while taller bars indicate a larger N400/P600 difference between conditions in the computational simulation. We can see that if we were to define the presence of an N400/P600 effect via a simple threshold, it would be possible to choose a threshold based on which the model successfully predicts N400 effects from all nine target experiments, since we see substantially greater differences between conditions for Reversal-2, Animacy-2/-3, and Substitution-1/-2/-3. As for the P600, with a simple threshold the model would successfully simulate the P600 effect in eight of nine experiments, correctly predicting a P600 effect for Reversal-1, Animacy-1/-3, Substitution-1/-2, and Preposition-1—but incorrectly predicting a P600 effect for Animacy-2. We include horizontal dashed lines to indicate post-hoc thresholds that enable this separation between significant and non-significant human ERP effects.

To verify the patterns suggested by these figures, we test for significant effects of condition in the model outputs by fitting a linear mixed effects model for each experiment. The dependent measure is the simulated ERP amplitude. We set *condition* as a two-level contrast (experimental condition vs. control condition), and we include by-item slope and intercept as random effects. The results of this statistical analysis, shown in Table 8, indicate that the simulated ERP components

Table 8
A linear mixed effects model that compares simulated N400 and P600 in experimental and control conditions across experiments. *p < .05, **p < .01, ***p < .001.

	N400 effect		P600 effect	
	t-value	*p*	*t-value*	*p*
Reversal-1	.42	.67	4.91	<.001***
Reversal-2	2.21	.03*	.89	.37
Animacy-1	1.13	.26	12.28	<.001***
Animacy-2	17.26	<.001***	11.14	<.001***
Animacy-3	13.43	<.001***	1.99	.04*
Substitution-1	2.27	.03*	7.09	<.001***
Substitution-2	3.72	<.001***	2.36	.02*
Substitution-3	4.25	<.001***	.19	.85
Preposition-1	0	1	10.59	<.001***

for which the model produces a significant effect are precisely those that fall above the post-hoc thresholds drawn in Fig. 7.

In summary, our model is able to replicate fully the patterns of N400/P600 effects for eight of our nine simulated experiments, including the divergent results between the two role reversal experiments characterized by identical linguistic manipulations, and the biphasic effects in Animacy-3, Substitution-1 and Substitution-2. The model is also able to capture the traditional syntactic P600 effect reflected in Preposition-1. The one failure of the model is its incorrect prediction of a P600 effect in Animacy-2—we will discuss the reasons for this incorrect patterning below. We now turn to a more detailed analysis of the dynamics within the model that enable it to produce these various patterns of effects.

6.1. Explanation of model patterns

We will first break down how the functioning of the model enables the eight successful simulations, before discussing explanations for the behavior on the P600 in Animacy-2.

Heuristic interpretation selection. The component of the model that plays an obvious role in these patterns is the heuristic interpretation mechanism: this is what enables the model to select an interpretation that deviates from the literal input, and this is also the component that aligns most closely with hypothesized processes driving existing theories of the semantic P600. As in verbally-specified theories, our model builds on the intuition that if the processor is entertaining a less anomalous interpretation of the input, this can lead to a reduced N400 in the anomalous condition and a reduced or eliminated N400 effect between conditions—while the need to reconcile this alternative interpretation with a more literal interpretation can lead to increased P600 in the anomalous condition and thus a P600 effect. When we examine the model's behavior in the simulations, we see that this is indeed the basic contribution of the heuristic interpretation mechanism: in general, when the model engages in error correction (selection of a less anomalous, non-literal interpretation), this serves to reduce N400 amplitude and increase P600 amplitude. However, there are two key differences that allow our model to account for more complex variations between experiments. First, in our model this heuristic interpretation process is probabilistic: rather than assuming that *all* anomalous sentences in the experimental condition will undergo error correction, our model estimates for each stimulus whether it will receive a revised interpretation, such that some experiments will have higher rates of error correction of their anomalous sentences than others (also in line with Ryskin et al. (2021)). Second, in addition to predicting whether a given item will error-correct, we use item-specific measures to quantify exactly how well the target word fits within the chosen interpretation (for the N400), and exactly how dramatic the reconciliation is between interpretations (for the P600). When we analyze the dynamics between all of these components, we find that the heuristic interpretation mechanism is indeed a significant contributor to simulated patterns—however, the interactions between rate of error correction, magnitude of target word probability in context, and magnitude of semantic divergence between interpretations give rise to a more complex overall picture which helps to account for variations between experiments that are challenging to explain with an error correction mechanism alone.

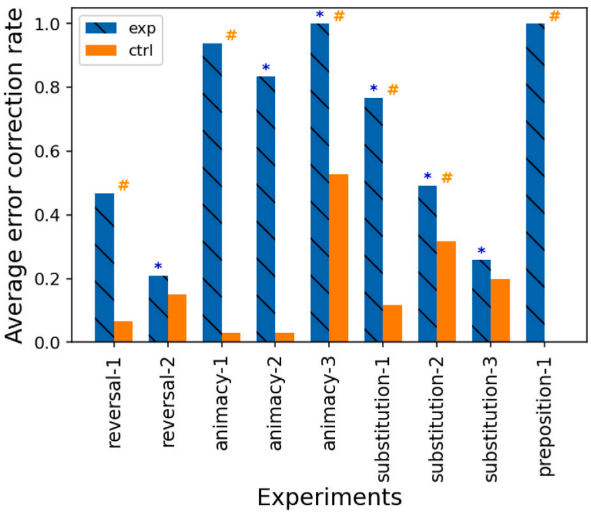
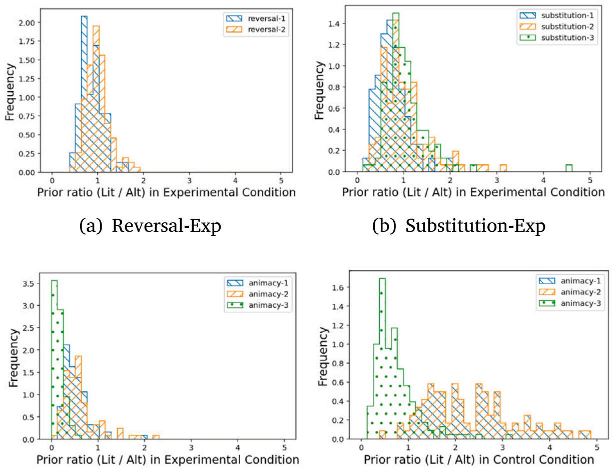


Fig. 8. Error correction rates for experimental and control conditions across experiments. Blue * and yellow # represent significant N400 effect and P600 effect, respectively, in original human experiments. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The heuristic interpretation process is driven by the noisy channel computation, in which the prior (inverse perplexity) and likelihood (string distance similarity) jointly determine whether to perform error correction. Since the likelihood of transformation between input and alternative sentences is roughly equivalent across experiments within our simulations, the most important factor driving differences in error correction rates between experiments is the prior, which is based in sentence probability and which we associate with plausibility. Specifically, we can trace the rate of error correction primarily to the distribution of differences in this plausibility prior between literal and alternative interpretations of items in each condition. In the experimental condition, when the more plausible alternative interpretation has a much higher prior (is much more probable) than the anomalous original sentence, error correction is more likely. Alternatively, if the priors of literal and alternative sentences tend to be very similar, then the cost of transformation (captured in the likelihood) is more likely to prevent error correction. For the purpose of examining model dynamics in this section, we define *prior ratio* of a given stimulus as $\frac{\text{Prior(Literal)}}{\text{Prior(Alternative)}}$. A low prior ratio would encourage error correction of that item into the alternative interpretation (because the literal interpretation has a much lower prior, and is thus marked as substantially less plausible, than the alternative).

Fig. 8 shows the model's error correction rates per condition for each of the simulated experiments, while Fig. 9 shows the distributions of prior ratios. We will refer to these figures in the analysis below.

Role reversal experiments. We will begin by discussing the model's behavior in the role reversal experiments. We see above that the model successfully predicts the divergent results between these experiments—that is, a P600-only effect for Reversal-1, and an N400-only effect for Reversal-2. What exactly in the models' mechanisms gives rise to this result? We see in Fig. 8 that the error correction rates in Reversal-1 differ much more between conditions than in Reversal-2. We can trace this back to the prior ratios shown in Fig. 9, where we see that the prior ratios in Reversal-1 trend lower than those in Reversal-2, with ratios consistently below 1 for Reversal-1, and ratios often exceeding 1



(a) Reversal-Exp

(b) Substitution-Exp

(c) Animacy-Exp

(d) Animacy-Ctrl

Fig. 9. The distributions of prior ratio between literal and alternative interpretations in experimental conditions for all experiments, and in control condition for Animacy experiments (Animacy-2 and −3 are those that contain alternatives that are not the same as items in the other condition).

in Reversal-2. This means that the model rates role-reversed sentences as substantially less good than the canonical sentences in Reversal-1, but it finds the role-reversed and canonical sentences to be of similar goodness in Reversal-2—with role-reversed sentences even being rated better in many cases. As a result, in Reversal-1 the model frequently error-corrects the role-reversed sentences to canonical sentences, and rarely error-corrects in the other direction. In Reversal-2, however, Fig. 8 shows that the error-correction rate is lower for role-reversed sentences, with the model error-correcting the control sentences to role-reversed sentences nearly as often.

These differences in error-correction patterns give rise directly to the differences in N400 and P600 effects. In Reversal-1, we have high rates of error correction of role-reversed to canonical sentences, and low rates of correction in canonical sentences, which means that a high percentage of trials will have the canonical sentences as heuristic interpretations for both experimental and control conditions. This leads to comparable N400 amplitudes between conditions, and thus no N400 effect. Additionally, since there is a large difference in error-correction between conditions – many heuristic interpretations in the experimental condition deviate from the literal syntactic interpretation, while most heuristic interpretations in the control condition do not – this predicts a difference in P600 amplitudes between conditions, and thus a P600 effect.

In Reversal-2, the error-correction rates for both conditions are fairly low, and the rates are more similar between conditions. This means that the model is less likely to process role-reversals with heuristic interpretations that match the canonical sentences, so a higher percentage of experimental trials will have larger N400 amplitude, while most control trials will have the lower N400 amplitude associated with canonical sentences (even though a small percentage of those trials in fact error correct to role-reversed interpretations). This difference between conditions leads to an N400 effect. Additionally, because the error-correction rates are much more similar between experimental and control conditions in Reversal-2, this indicates that the conflict resolution costs will be comparable between conditions, resulting in no P600 effect.

It is worth acknowledging at this point that in our model it is not exclusively items in the experimental condition that undergo "error correction"—a small percentage of control items are in fact corrected to alternatives that we would by default consider *less* plausible. As we discuss below, to an extent this should probably be considered noise in the neural network measures. However, we do not discount the possibility that control items may at times also receive non-literal heuristic interpretations in reality. We will explore this topic in more detail in Section 7.

Animacy experiments. For Animacy-1 and Animacy-3, the model successfully simulates P600-only and biphasic effects, respectively, and the model correctly predicts an N400 effect for Animacy-2, though it incorrectly predicts a P600 effect for Animacy-2. Here we discuss how these different patterns arise.

For Animacy-1, in Fig. 8 we see that the difference in error correction rates between conditions is extremely large, as a result of very low prior ratios in the experimental condition, mirrored by very high prior ratios in the control condition, as shown in Fig. 9. This means that the model finds the plausible alternatives much better than the literal interpretations of the experimental sentences, leading to a high error correction rate. As in the role reversal experiments, the error-corrected alternatives for the experimental sentences are the same as the corresponding control sentences, so the high rates of error correction in the experimental condition will lead to the N400 being computed on similar sentences in both conditions, leading to no N400 effect. Additionally, due to large difference in error correction rates between conditions, there will be more significant reconciliation between interpretations in the experimental condition than in the control condition, so we see a P600 effect. The reasons for these patterns largely mirror those in Reversal-1, but the differences are even larger than in Reversal-1.

For Animacy-3, the prior ratios in the experimental condition trend very low, which again leads to a high error correction rate in the experimental condition (100%). Incidentally, the prior ratios in the control condition also drive a fairly high proportion of control sentences to be reinterpreted as alternatives (52.8%), though there remains a large difference between conditions in error correction rates. Just as in the above experiments, the larger error correction rate in the experimental condition than in the control condition gives rise to the P600 effect—there is more reconciliation to be done in the experimental condition than in the control condition, and this generates a P600 effect. So how do we generate an N400 effect, despite the fact that all of the experimental sentences are correcting to more plausible alternatives? To understand why, it is important to remember that for Animacy-3 (unlike the three experiments analyzed so far) when we error-correct the experimental sentences to their more plausible passive alternatives, these plausible alternatives are *not* the same as the corresponding control sentences. So although the experimental sentences are error-correcting, this does not have the same effect as in the prior experiments (eliminating the N400 effect due to similar sentences in both conditions). Instead, we find that the conditional probability of the target in the more plausible error-corrected experimental sentence is *still* less probable than either version (literal or error-corrected) of the control sentence. For example, for the experimental sentence *The journalist astonished the article*, the model is correcting to the more plausible alternative *The journalist was astonished by the article*—but the probability of *article* in this sentence is still lower than either the passive or the active version of the corresponding control sentence (Literal: *The journalist wrote the article*; Alternative: *The journalist was written by the article*). We can see these patterns at an aggregate level in Fig. 10.

This means that for this experiment, the N400 effect can arise independent of error correction rates, because the closest alternatives for the experimental sentences still yield less good target word fit than the control sentences.

What this finding highlights is that we can increase the power of our theories to account for biphasic effects by simply allowing for the possibility that the processor corrects to more plausible non-literal interpretations which nonetheless are still less plausible than the control sentences. This finding also indicates that our manual choice of alternative interpretation makes a significant contribution to our

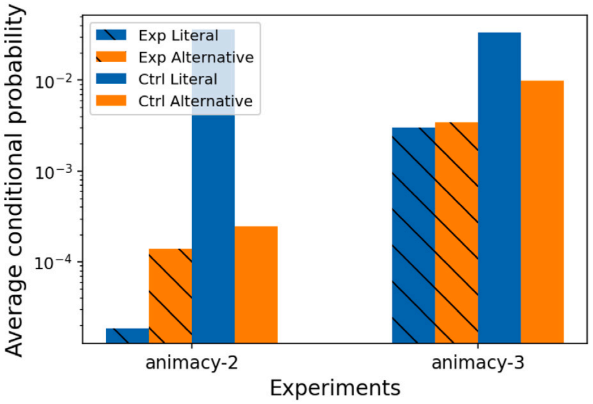


Fig. 10. Average conditional probability (log-scale) for literal and alternative interpretations of experimental and control conditions in Animacy-2 and Animacy-3.

ability to simulate the ERP patterns in this experiment. It is important to reiterate, therefore, that we chose these alternatives based on a simple principle of identifying a minimal edit that would result in a more plausible interpretation. While we cannot guarantee that our selected alternative interpretations are precisely those entertained by the processor, the success of the simulation supports the possibility that the processor may be doing error corrections resembling these.

Finally, Animacy-2 is the experiment for which our model correctly predicts an N400 effect, but incorrectly predicts an additional P600 effect. Similarly to Animacy-3, the prior ratios in Animacy-2 trend very low in the experimental condition, leading to a high error correction rate—and also similarly to Animacy-3, the conditional probabilities for both the literal and alternative interpretations in the experimental condition are lower than interpretations in the control condition. This leads Animacy-2 to show a biphasic effect for similar reasons to Animacy-3: although a significant percentage of sentences are error corrected, we still produce an N400 effect because the targets in the error-corrected sentences remain less probable than in the candidate interpretations in the control condition. While this results in the correct pattern for the N400 component, we are inclined to believe that in this case the successful N400 simulation is a coincidence—as we will discuss below, we believe that the model's treatment of passives leads to a divergence from human processing, generating a high error correction rate when this may not cognitively be the case.

Substitution experiments. The model also successfully simulates biphasic N400-P600 effects for Substitution-1 and Substitution-2, and an N400-only effect for Substitution-3.

For Substitution-1, we have again a large difference in error-correction rates between experimental and control conditions, driven by generally low prior ratios in the experimental condition (mirrored by high prior ratios in the control condition). As in the above experiments, the difference in error correction rates between conditions generates a straightforward P600 effect. How, then, is the N400 effect generated? In many ways the patterns here resemble Reversal-1, in that there is much more error-correction in the experimental than the control condition, and the error-corrected versions of the experimental sentences are again the same as the control sentences—so we may expect absence of an N400 effect, since the heuristic interpretations for a majority of items will be matched across conditions. However, upon closer examination we find that the items from this experiment differ on another measure: the conditional probabilities, shown in Fig. 4, which drive the simulated amplitude of the N400. We see in that figure that Substitution-1 has a much larger difference in conditional probabilities between experimental and control sentences, by comparison to Reversal-1—this means that the model finds the target words in the uncorrected experimental sentences particularly surprising in Substitution-1. The consequence is that even with a rather small percentage of experimental condition items that remain uncorrected (24.4%), the much larger N400 amplitudes from these uncorrected items are strong enough to generate an overall N400 effect relative to the control condition.

For Substitution-2 the model again succeeds in generating a biphasic effect, as in Substitution-1—but in Substitution-2 this effect occurs for slightly different reasons. The experimental condition error correction rate is lower than Substitution-1, which leads to a more straightforward N400 effect: for most trials, the N400 operates on the literal, anomalous experimental sentence, so the N400 amplitude is larger in the experimental condition. The question, then, is how the P600 effect is generated, given that error-correction rates between experimental and control conditions are similar—after all, in Reversal-2 we saw that similar error-correction rates led to a lack of P600 effect. For Substitution-2, we find the explanation in the patterns of cosine similarity between literal and heuristic interpretations, which is used to simulate P600 amplitude—the lower the cosine similarity, the higher the P600 amplitude. In Fig. 5, we see that average cosine similarity between literal and alternative interpretations is much lower for Substitution-2 than for Reversal-2. This means that the model considers error corrections in Substitution-2 to involve more dramatic semantic deviation, so the reconciliation process generates stronger P600 amplitudes. As a result, the relatively small difference in error correction is amplified into a large difference in P600 amplitudes, and a resulting P600 effect.

For Substitution-3 we see error correction patterns that are again similar to Reversal-2 (even a bit more so, since the difference in error correction rates between conditions for Substitution-3 is slightly smaller than Substitution-2). We see an N400 effect for the same reasons as Reversal-2 and Substitution-2: relatively low error correction overall, leading to less plausible heuristic interpretations in the experimental condition, higher N400 amplitudes in that condition, and an N400 effect between conditions. As for the P600, due to the small difference in error correction rates between conditions, we expect the P600 amplitudes to be similar between conditions as well. Unlike Substitution-2, Fig. 4 shows that the average cosine similarity for Substitution-3 is high, comparable with Reversal-2—so the small difference in error correction rates, combined with the relatively high cosine similarities, leads to a lack of P600 effect.

Preposition experiment. Finally, the model also successfully simulates the P600 effect elicited by syntactic anomalies in Preposition-1. The explanation of the model's P600-only effect here is straightforward: as we see in Fig. 8, all sentences in the experimental condition are error corrected by the model, while none of the control sentences receive error correction. This error correction pattern is driven by the large difference in probabilities that the neural network assigns to the grammatical versus ungrammatical alternatives, as seen in Fig. 2. Because of the large difference in error correction rates between conditions, there is a straightforward P600 effect—and because the ungrammatical experimental sentences are all being corrected to their grammatical counterparts, which match the corresponding sentences in the control condition, the N400 acts on the same heuristic interpretations in both conditions, leading to a lack of N400 effect.

Unsuccessful simulation: Animacy-2 P600. The primary failure in these simulations is that the model simulates a significant P600 effect for Animacy-2, while the human experimental results show no P600 effect. Why does this happen, and what should we conclude from this incorrect pattern in the model?

The stimuli in Animacy-2 are designed such that the subject in the experimental sentence should not be thematically attracted to the main verb (e.g., *The dusty tabletop was devouring...*), so the obvious alternative interpretation (e.g., *The dusty tabletop was devoured...*) is not semantically more plausible than the original. The intention is that the low plausibility of the alternative should block error correction, such that there should be a classic N400 effect with no P600 effect. We have seen our model generate this type of effect with Reversal-2 and Substitution-3. However, in the case of Animacy-2, our model priors fail to reflect the fact that the alternative interpretations in the experimental condition should not be consistently better than the literal interpretations: we see in Fig. 9 that the prior ratios trend very low for the experimental condition—and more so than in the control condition—such that the model finds the alternative interpretations substantially more tempting in the experimental condition, leading to higher error correction and a P600 effect between conditions.

What this means is that the model is consistently preferring the error-corrected alternative of the experimental sentences, despite the fact that these stimuli are designed to block error correction—and the lack of a P600 effect in humans suggests that this blocking is successful. So why would the model prefer *The dusty tabletop was devoured* over *The dusty tabletop was devouring*, despite the fact that neither sentence is plausible? We suspect that the reason for this preference is that the model has learned that inanimate subjects are unlikely to be the subjects of active verbs, so despite the residual semantic strangeness, the model prefers the passive version on the basis of the improved syntactic compatibility. This suggests a limitation in alignment between the plausibility mechanisms driving error correction in the brain and the neural-network-based perplexity measure used here: the Animacy-2 human results suggest that improvement on this syntactic dimension is not sufficient to drive an interpretation shift in humans, while it is enough to do so in the model. Notably, though error correction on this particular syntactic basis seems misplaced, the model's inclination toward error correction based on ungrammaticality is instrumental to our successful Preposition-1 simulation—so better aligning with the human mechanism is likely more complex than removing syntactic sensitivities entirely. Note that the model does still generally rate the experimental sentences as worse than the control sentences, both before and after error correction—so the model is to an extent sensitive to the strangeness of both *The dusty tabletops were devouring* and *The dusty tabletops were devoured* by comparison to *The hearty meal was devouring*. However, the models diverge from human patterns in preferring the passive version when the inanimate entity is the subject.

This brings us back to our discussion above of the model's correct prediction of an N400 effect in the Animacy-2 simulation. The model produces an N400 effect despite the high rate of error correction, and this is because the error-corrected interpretations are still rated as less plausible than the control sentence interpretations (as in Animacy-3). However, we have just reasoned that the model's high rate of error correction on these sentences is a deviation from the behavior of cognitive mechanisms—preferably, the model should adjust its plausibilities such that the syntactic improvement in the passive sentences is not enough to prompt error correction in the first place. So while the N400 effect in our simulation does align with the human results, we suspect that the more cognitively plausible scenario would be for the N400 effect to be generated on the basis of non-error-corrected stimuli in this case. In Section 7 we discuss possible means of addressing this divergence between model and human behavior.

Summary. We see in these analyses that, in keeping with traditional theories, a central driving influence on ERP patterns in our model is whether the stimuli are error-corrected into more plausible alternatives during the heuristic interpretation stage. Error correction in general decreases the amplitude of the N400 by making the interpretation more plausible, and increases the P600 by increasing semantic divergence between interpretations to be reconciled. However, our model reveals a more complex picture that helps to account for a greater amount of the variation observed in human results. First, we see that in some cases, if the error-corrected interpretations of anomalous experimental sentences are still more surprising than the control sentences, then high rates of error correction will not erase the N400 effect, enabling biphasic effects. Second, we see that if conditional probabilities of target words in context are dramatically different between conditions, we can again produce an N400 effect even when the error correction rate is relatively high—because even a small number of uncorrected sentences can produce an inflated average N400 amplitude. Similarly, for the P600 component, in cases where there is particularly large semantic divergence between heuristic and true interpretations, this can result in P600 effects even in experiments with relatively low rates of error correction.

These simulations allow us to tease apart how dynamics of individual stimulus properties can help to explain the observed effects, within frameworks firmly rooted in existing theory, but elaborated through our model. Below we will discuss more about what we learn from our model's behavior, and what limitations and future work remain.

7. Discussion

We have presented a computational model of the mechanisms underlying the N400 and P600 response patterns from a range of psycholinguistic studies. Our model is founded on execution of a probabilistic early interpretation stage, during which a heuristic candidate interpretation is selected on the basis of a noisy-channel computation. This interpretation selection reflects a trade-off between the plausibility of the candidate interpretation—as estimated by sentence probability from a pre-trained neural network—and amount of revision relative to the literal interpretation. The selected interpretation then plays a key role in both N400 and P600 components: N400 amplitude reflects the fit of the target word in the context of the selected heuristic interpretation, and P600 amplitude reflects the semantic reconciliation between the heuristic and literal interpretations.

This model is founded on many of the same insights as existing theories—in particular, the notion that sentence processing involves a plausibility-driven interpretation component, which may yield interpretations that differ from the literal interpretation. However, rather than assuming fairly uniform behavior across stimuli in a given experiment, our model makes fine-grained estimates of idiosyncratic statistical properties of each stimulus, and of the corresponding impacts on the hypothesized mechanisms driving ERP components. These estimates come with certain tradeoffs, as we discuss further below—however, as a result of capturing this item-level variation, we are able to account for patterns in human results that have been challenging to explain with verbally-specified theories, including divergent effects for ostensibly identical role reversal experiments, and biphasic effects that have presented challenges for classic heuristic interpretation accounts.

In incorporating fine-grained stimulus-level variation, our model takes a step beyond computational models that have simulated N400 and P600 phenomena using synthetic data and manually-engineered environmental statistics (Brouwer et al., 2017, 2021; Rabovsky et al., 2018). The importance of capturing idiosyncratic stimulus properties is supported by substantial experimental evidence suggesting that ERP responses are sensitive to a wide range of linguistic factors, including frequency of words (Dufour, Brunellière, & Frauenfelder, 2013), semantic association between targets and preceding contexts (DeLong & Kutas, 2020), degree of contextual constraint (Federmeier, Wlotko, De Ochoa-Dewald, & Kutas, 2007), sentence plausibility (Van Petten & Luka, 2012), and predictability of targets (DeLong, Quante, & Kutas, 2014). In ERP experiments, it is almost impossible to create stimuli that control all variables that may exert influence on experimental outcomes. As a result, theories that focus exclusively on effects of a small set of linguistic variables of interest risk missing interactions with other important variables—and as we discuss in Section 2, similar limitations apply to computational models that use synthetic stimuli generated based on only a small number of target variables. Our model, by contrast, runs on real experimental stimuli and incorporates estimates from neural networks trained on natural data in order to approximate the effects of such idiosyncratic variation.

At the same time, our model also takes a step beyond computational simulations that have used neural network measures trained on natural data, but that use these measures largely in isolation to simulate either N400 or P600 components. As we discuss in Section 2, these studies have shown clear evidence of alignment between these measures and ERP components, but evidence also indicates that these measures alone are not enough to account for the full scope of patterns observed in the combined N400 and P600 components that we aim to account for here. Our model takes advantage of the estimates provided by these measures, but incorporates them within subcomponents of a larger model designed to test psycholinguistic hypotheses in a more targeted manner.

The success of our model lends stronger credence to the basic principles driving heuristic interpretation theories, while further elaborating on several details of the relevant mechanisms in ways that strengthen the explanatory coverage of the theories. Our model posits that the processor employs a rational inference mechanism to select heuristic interpretations. This rational inference process takes the form of a noisy channel model, which ranks interpretations based on a tradeoff between plausibility and extent of revision. Our account posits that the driver of N400 amplitude is fit of target word (indexed by conditional probability) in the context of this selected heuristic interpretation, while the driver of P600 amplitude is the extent of reconciliation needed between the heuristic interpretation and literal interpretation (indexed by semantic divergence). We are not currently making any strong assumptions about whether the heuristic interpretation is a temporary interpretation that gives way to the literal, or a candidate interpretation that persists throughout the comprehension process. Our model does, however, assume that by the time the P600 is generated, the processor has access to both the literal interpretation and the heuristic interpretation, since the reconciliation between these interpretations drives the P600 amplitude. Additionally, our model offers a number of further theoretical details involving the role of fine-grained stimulus variation, and how these stimulus properties interact with the posited processing mechanisms to produce the observed ERP patterns. In particular, our model indicates that error correction is only one contributor to presence or absence of N400 and P600 effects: if error correction yields interpretations that are still less plausible than control sentences, or if items are characterized by extreme levels of target word conditional probability or divergence between interpretations, then these factors can explain patterns in the ERP components that are not predicted by the error correction process alone.

Use of neural network measures. For capturing stimulus-level variation, our model relies on three primary estimates from neural network models: (1) estimate of the probability of an interpretation sentence, which we use to capture a notion of plausibility in the prior of our noisy channel model, (2) estimate of the conditional probability of the target word given the interpretation context, which we use to simulate N400 amplitude, and (3) estimate of the semantic divergence between the heuristic and literal interpretations, which we use to simulate the P600 amplitude. The success of our simulations not only provides support for the general mechanisms posited in our model, but also supports the capacity of these neural network measures to serve as reasonable proxies to predict the specific behaviors and sensitivities of these mechanisms.

While these neural network measures confer non-trivial benefits in the model, it is also important to acknowledge limitations that come with our use of these measures. The first limitation is in transparency: while we can confidently characterize these estimates as involving probability of a sentence, probability of a word in context, and divergence between sentence representations respectively, the precise linguistic components that drive these estimates remain relatively opaque. This means that although the model on the whole is comparatively transparent, for the subcomponents of the model defined by these estimates we cannot make strong claims about the specific linguistic cues that drive the corresponding behaviors.

One consequence of this is that we deviate from previous verbally-specified theories that have focused on the specific balance of linguistic cues driving the heuristic interpretations. Instead, in our selection of the heuristic interpretation we use a fairly general notion of plausibility estimated by sentence probability, and this probability measure incorporates a variety of cues, the types of which are not fully specified. This general notion of probability applies also to our conditional probability measure for N400 amplitude. As we have noted above, these neural network sentence probabilities and target word conditional probabilities do differ from human offline plausibility judgments and cloze probabilities, respectively. One of the most salient such divergences is in the models' inclination to assign higher probability to anomalous interpretations that are semantically or syntactically similar to plausible interpretations (LeBrun et al., 2021; Michaelov & Bergen, 2020). Additionally, the neural network sentence probabilities exhibit more sensitivity to ungrammaticality than would be predicted by an intuitive notion of plausibility per se. In this sense, our model's use of these measures is positing that the notion of plausibility that influences the heuristic interpretation process, and the notion of target word fit that drives N400 amplitude, are slightly broader, more statistically driven notions of plausibility and fit than are embodied in human offline plausibility judgments and cloze probabilities. We note that our use of probabilistic measures of this kind also aligns with the observation that neural network conditional probabilities predict N400 amplitudes better when the probabilities highly correlate with semantic similarity between context and target (Michaelov et al., 2021).

As for the semantic divergence measure that drives P600 amplitude—this measure too should be interpreted with some care. We compute this divergence between sentences using representations from a model trained on a semantic task, so as to maximize the semantic nature of the representations and their divergence estimates. However, it is important to acknowledge that the exact information encoded in these sentence representations remains relatively opaque, and their ability to capture complex semantics is likely still limited (Ettinger, 2020; McCoy, Min, & Linzen, 2020; Min, McCoy, Das, Pitler, & Linzen, 2020; Yu & Ettinger, 2020). This means that our model currently simulates the P600 as being driven by a relatively coarse-grained notion of semantic divergence between literal and heuristic interpretations. As we mention above, it is also possible that the semantic focus of our divergence measure may result in underestimation of P600 amplitudes in cases of syntactic anomaly—however, our measure shows enough sensitivity to syntactic divergence that it is able to produce a syntactic P600 effect, suggesting that to some extent our divergence measure also incorporates sensitivities outside the traditional scope of semantic similarity. It will be important in future work to investigate both the true semantic content of these neural network representations, as well as the possibility that the relevant divergence mechanisms are not strictly semantic.

The second limitation is that these neural network models have not originally been designed as psycholinguistic models – they are designed for engineering purposes in artificial intelligence – so what alignment they show with human mechanisms arises somewhat as a matter of coincidence, and not because the modelers intended to hypothesize and simulate particular human mechanisms. (Of course, the fact that these models are designed with the high-level goal of replicating the human capacity for language does make areas of alignment rather less coincidental.) Ultimately these neural network language models are simply large, sophisticated learning models that have high sensitivity to statistical properties of language, and that learn to encode many properties of language in their internal representations. The success of our model suggests that the statistical sensitivities and representational properties captured by these neural networks can provide reasonable estimates of the corresponding human properties that we wish to simulate in the relevant subcomponents of our model. However, there are bound to be divergences, and the clearest such case in our simulations is the case that leads to the unsuccessful Animacy-2 simulation: the neural network probabilities consistently favor passive sentences when the subject is inanimate, even if the passive sentence makes no more sense than the original. The fact that this results in a divergence from the human patterns suggests that mechanisms driving plausibility-based probabilistic inference in humans are less reliant on this particular syntactic dimension than are the probabilities produced by the neural network. Another area of potential divergence can be found in our model's prediction of error correction for a percentage of control sentences, resulting in correction to what we would expect to be *more* anomalous heuristic interpretations. It is important to consider the possibility that the mechanisms driving heuristic interpretation may indeed at times favor interpretations that are not favored based on humans' conscious ratings of plausibility, and that the probabilities assigned by the neural network models may better predict some such cases. However, it is likely that at least some of these cases represent additional divergences from human plausibility mechanisms influencing heuristic interpretation, marking further dimensions on which these probability measures can be refined for greater cognitive accuracy.

Further improving the alignment of the model with human patterns will require adjustment to the estimates that we are currently drawing from neural network models. Unfortunately, it is not trivial to make fine-grained changes to these estimates directly, since the neural network models develop their patterns based on generalized learning mechanisms operating on statistics in data. However, there are a couple of immediate possibilities worth testing. One possibility is to explore neural network models that achieve better performance on NLP benchmarks—better (typically meaning more humanlike) performance could potentially correspond to more humanlike plausibility behaviors. (It is worth noting, however, that we did test estimates from a number of highly successful neural network language models – BERT, RoBERTa, GPT-2 and GPT-3 – and all of these language models show the bias toward passive constructions that leads to the Animacy-2 divergence.) Another, perhaps more promising possibility would be to leverage fine-tuning processes, modifying the language models' sensitivities via different, more semantically-driven tasks such as labeling of thematic roles. Greater sensitivity to semantic dependency statistics may help to produce the distinctive Animacy-2 behavior that our model is currently missing: support for this possibility is provided by Michalon and Baggio (2019), who find that Animacy-1 and Animacy-2 have different profiles in terms of whether the target verb can be a direct dependent of the subject noun in the Wikipedia corpus. Notably, however, as we have observed above, improving alignment with human mechanisms will likely be more complex than simply making the measures strictly more "semantic", since our model benefits importantly from its syntactic sensitivities in simulating the syntactic P600.

If these more straightforward approaches are unable to improve the alignment with human mechanisms, then the next step will be to investigate alternative means of estimating probabilistic and representational properties altogether—in these simulations we have benefited greatly from the estimates extracted from pre-trained neural network language models, but these are by no means the only possible source of such probabilities and representations. More direct experimentation with different types of statistical learning models and representational models has the potential to allow finer-grained adjustment to these critical estimates in our model, while yielding more transparent and comprehensive insights into the posited cognitive mechanisms of plausibility, fit to context, and reconciliation between interpretations. This will be a valuable but substantial undertaking, which we leave to future work.

As a final note: in this paper we have highlighted the benefits of the neural network measures in allowing our model to capture complex interactions between individual stimuli and our posited mechanisms. Another valuable future direction will be to leverage the model's ability to generate item-level quantitative predictions, and analyze the extent of model alignment with human patterns at the item level. This will provide a still richer source of data for further refinement of the model's approximation of human mechanisms.

Candidate interpretations. An important simplification that we have made in our model is in the nature of candidate interpretations: in our simulations, the model is only selecting between one literal and one hand-picked alternative, when in reality the set of possible candidate interpretations will presumably be much larger and more complex. In selecting plausible alternatives for the anomalous sentences, we have endeavored to follow consistent principles and to choose what can reasonably be considered one of the closest possible plausible alternatives. As we discuss in Section 6.1, we cannot guarantee that our chosen alternatives align with those selected by the processor in reality—however, the success of the simulations indicates that this combination of hypothesized candidate interpretations and hypothesized processing mechanisms provides a potentially viable account for the observed ERP patterns in these experiments. Moreover, the dynamics that emerge with these interpretation candidates are also informative—in particular, we see that biphasic effects can straightforwardly be achieved if error correction of anomalous sentences yields interpretations in which targets are still less probable than in the canonical condition. Even if our chosen interpretations are not those that the processor entertains, this is a consideration that will be important in future modeling of these phenomena.

Another simplification in our candidate selection process is the fact that the inference process operates over fully-formed hypothesized interpretations, implying that the processor forms probability distributions over full candidate interpretations rather than filtering to subsets of cues. Additionally, our model selects a single heuristic interpretation based on the highest posterior probability, rather than maintaining a full probability distribution over possible constructions (Levy, 2008; Levy, Bicknell, Slattery, & Rayner, 2009). Though these simplifications allow us to make substantive progress in understanding how to account for observed ERP patterns, ultimately it will be important to model the full process of generating candidate interpretations given a literal sentence input. Modeling of this candidate generation process could incorporate more direct influence of theories of relevant cue types, and could make use of the full probability distribution over the interpretation space (Jurafsky, 2003; MacDonald, Pearlmutter, & Seidenberg, 1994; McClelland, 1986). Modeling of this process can also naturally engage with the long-standing debate between probabilistic serial models and probabilistic ranked parallel models (Gibson & Pearlmutter, 2000; Lewis, 2000).

Compatibility with broader N400 landscape. There is a robust history of debate about the relationship between the N400 and predictive processing, and use of predictive metrics to approximate the N400. Theories of the N400 have in particular debated whether the N400 reflects context-based pre-activation prior to arrival of upcoming words (Cheimariou, Farmer, & Gordon, 2019; DeLong, Urbach, & Kutas, 2005; Kutas & Federmeier, 2000; Szewczyk & Schriefers, 2018; Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005; Wicha, Bates, Moreno, & Kutas, 2003), semantic integration of current input into previous representations of context (Hagoort, Baggio, & Willems, 2009), or a combination of distinct cognitive processes, such as an interaction of both pre-activation and integration (Calloway & Perfetti, 2017; Nieuwland, Barr, Bartolozzi, Busch-Moreno, Darley, Donaldson, et al., 2020; Van Berkum, 2009). We argue that simulating the N400 via word probabilities conditioned on context, as we do in our model, is in principle compatible with any of these views—these probabilities reflect contextually-conditioned expectations for upcoming words, but they can also serve as proxies for how well a word fits with (and by extension how well it can be integrated into) a context. Note that we use neural network language model probability here, but use of conditional word probability for the N400 can accommodate a wide range of probabilistic models with the potential to reflect broad variation in underlying mechanisms.

What about N400 patterning beyond that in the studies simulated above? Our simulations here have focused primarily on accounting for semantic anomaly patterns by incorporating a heuristic interpretation process, such that the N400 reflects conditional probability of a word within a potentially non-literal interpretation of the context. However, as we have discussed in Section 2, broader N400 patterns can at times be explained via conditional probabilities from neural network language models alone, without intervention by a heuristic interpretation mechanism. From the perspective of our modeling framework, when neural network probabilities appear to explain N400 patterning directly, we may assume that the heuristic interpretation mechanism is yielding low rates of non-literal interpretation, such that the heuristic interpretation matches the literal interpretation, and the N400 behaves as it would if the heuristic interpretation stage did not exist. Future work can determine whether experiments in which neural network probabilities better explain N400 patterns are also those with lower rates of error correction to non-literal interpretations.

While we can explain high correlation between N400 and conditional probability in terms of low error correction rate, it is also worth acknowledging that these neural network conditional probabilities can at times show behaviors not unlike those accounted for by the heuristic interpretation process—for instance, assigning high probability to certain verbs in the role reversal paradigm regardless of the role configuration of the nouns (Ettinger, 2020; Lindborg & Rabovsky, 2021; Michaelov & Bergen, 2020). This raises the question of whether these neural network models may have already developed certain strategies that overlap in functionality with our heuristic interpretation mechanism. To address this possibility, it is necessary that we continue to work to increase the transparency of these neural network measures, and to clarify the details of their alignment and divergence from human patterns. As we come to better understand the contributing factors driving the conditional probabilities from these models, and the extent to which their alignment with N400 patterns are indeed satisfactorily explained based on error correction rates from our posited heuristic interpretation mechanism, we will be able to clarify further whether there are redundancies to be addressed between these mechanisms.

Compatibility with broader P600 landscape. Beyond discussions of multi-stream and single-stream models, as described in Section 2, the P600 literature has also involved important debate about whether P600 amplitude reflects syntactic reanalysis (Kim & Osterhout, 2005), effort of integrating structural information (Bornkessel-Schlesewsky & Schlesewsky, 2008; Kaan, Harris, Gibson, & Holcomb, 2000) or monitoring the size of errors (Kolk et al., 2003; Van de Meerendonk, Indefrey, Chwilla, & Kolk, 2011; Van Herten et al., 2006, 2005). In particular, error monitoring theories can be linked to theories that view the P600 as a subcomponent of the more domain-general P300 signal, which is responsive to low-probability events. The latter theories suggest that syntactic and semantic P600 may be separate components entirely (Leckey & Federmeier, 2020). These functional and mechanistic debates about the P600 are largely orthogonal to the multi-stream/single-stream model debate.

As we claim in Section 2, our model is largely compatible with either a multi-stream or a single-stream framework: one could assume our heuristic interpretation process to reflect a separate mechanism or stream from that which produces the literal interpretation, or one could imagine the literal interpretation to emerge as a later interpretation within the same processing stream. As for the reanalysis/integration debate, our model also largely abstracts away from the particulars of this discussion, in that we simulate P600 amplitude by quantifying divergence between heuristic and literal interpretations, but do not commit to the specific mechanisms used to reconcile between these interpretations. This measure of divergence may superficially be more aligned with a reanalysis mechanism, in that it involves calculating divergence between updated and previous interpretations—however, one can nearly as easily imagine that this measure of divergence could quantify difficulty of integration, if the heuristic interpretation reflects existing structure and the literal interpretation reflects new information to be integrated. Our model is most obviously compatible with P300 theories of the P600, where the P600 is manipulated by errors—however, this class of theory diverges from our noisy-channel framework in not considering whether the errors are likely to be corrected. Additionally, while our model does not rule out the possibility that semantic and syntactic P600 are separate components, the capacity of our model to account for both types within a single framework suggests the potential for shared mechanisms underlying both.

How feasible is it to extend our model to account for the broader P600 literature? We have already seen that our model can account for semantic P600 patterns as well as traditional P600 effects in response to syntactic errors. Our model predicts that the P600 to syntactic and non-syntactic violations will operate comparably: the P600 amplitude will be decided by (1) whether there is an error correction and (2) how dramatically the corrected interpretation differs from the literal interpretation. One straightforward extension would be to account for P600 effects in response to spelling errors (Van de Meerendonk et al., 2011; Vissers, Chwilla, & Kolk, 2006). Similarly to our Preposition-1 simulation, we can expect that our model will produce high rates of error correction in the case of minor edits required to fix spelling errors—and this will in turn force reconciliation between divergent literal and heuristic interpretations, leading to larger P600 amplitudes in erroneous sentences. Along the same line, the model could also be applied to explain P600 effects to garden-path sentences (Frisch, Schlesewsky, Saddy, & Alpermann, 2002; Gouvea, Phillips, Kazanina, & Poeppel, 2010; Kaan & Swaab, 2003; Osterhout & Holcomb, 1992; Osterhout, Holcomb, & Swinney, 1994): when presented with low-probability garden-path sentences, the model could be expected to select a structurally-simpler alternative heuristic interpretation, and the divergence between this alternative and the literal interpretation would drive an increased P600. Given that different interpretations of garden-path sentences share the same set of surface forms, expanding to this type of study would likely involve extending prior and likelihood functions to operate directly on syntactic structures.

Our model could also be extended to account for the influence of plausibility rate and task demands on the P600. The semantic P600 has been reported to be sensitive to proportion of problematic sentences (Gunter, Stowe, & Mulder, 1997; Hahne & Friederici, 1999) and nature of task demands (Zwaan & Radvansky, 1998)—specifically, P600 effects are reduced when there is a higher rate of anomalous items, and P600 effects are larger when tasks encourage comprehension or evaluation of a sentence. We anticipate that we can account for such effects by assuming that stimulus properties and task demands within experimental contexts can temporarily affect estimates of the prior within the noisy channel model. For instance, if there is a large proportion of implausible events being described, comprehenders may be more likely to regard anomalous sentences as acceptable (as, for instance, when reading scientific fiction). In this case, the prior $P(m)$ may come to assign higher probability to utterances that are less likely in the real world, reducing error correction—and by extension, the P600. The capacity of a noisy channel model to capture this type of effect has already been demonstrated by Gibson et al. (2013), who use a noisy-channel model to show that when larger proportions of semantically implausible sentences cause a shift in the model prior, this leads comprehenders to be less likely to interpret sentences as plausible alternatives in offline sentence comprehension tasks. As for the effect of task demands, we anticipate that this can be accounted for by adjusting the levels of linguistic information that impact the prior: when doing a comprehension task, the prior would operate as assumed in the current model—but when doing a more superficial task, human "plausibility" estimates may be driven by more superficial factors such as syntactic acceptability. Accounting for this in our model could be done fairly straightforwardly by manipulating the prior, but would require exerting more control over the prior estimate than we currently do with neural network perplexities. As we discuss above, more controlled modification of the estimates that we currently draw from neural networks will be a non-trivial but important direction for future work, as this will allow for greater transparency in adjudicating between contributions of different linguistic cues, while also enabling finer-grained improvements in the accuracy of our model and its application to still broader sets of phenomena.

8. Conclusion

In this paper, we present a computational model to account for patterns of N400 and P600 ERP components in response to semantic violations (as well as syntactic violations), simulating nine studies featuring N400-only, P600-only, and biphasic N400-P600 effects. Our model builds on the foundation of existing psycholinguistic theories that posit what we refer to as *heuristic interpretations*, which can diverge from the literal interpretation of the input. In our model we formulate the heuristic interpretation process as a probabilistic selection among candidate interpretations driven by a noisy-channel model computation. N400 amplitude is then simulated based on the probability of the target word in the context of the heuristic interpretation, representing sensitivity to fit of target word in context—and P600 amplitude is simulated based on semantic divergence between heuristic and literal interpretations, representing reconciliation between these interpretations. In our simulations we use the real experimental stimuli from the human experiments, and leverage measures from neural network models trained on large amounts of data in order to capture fine-grained variation in stimulus properties. As a result, we are successfully able to reproduce patterns of N400 and P600 effects for eight of the nine studies that we simulate, including accounting for challenging patterns such as divergent effects for identical role reversal manipulations, as well as biphasic effects. The model's behaviors shed light on more nuanced potential explanations for observed ERP patterns, in which heuristic interpretation processes are only part of a larger story, and finer-grained probabilistic and representational properties of stimuli have additional non-trivial influence on observed effects. In the case of the single study that the model does not successfully simulate, we are able to identify a straightforward explanation in terms of divergence between what we infer to be factors driving human plausibility mechanisms, and factors influencing our plausibility proxy drawn from the neural network model. On the whole, the success of the simulations indicates that the theory embodied by our model represents a strong candidate account for mechanisms driving the observed ERP patterns. Further work can continue to investigate the breadth and granularity of results that the model is able to account for, continue to refine the quantitative measures that we currently draw from neural network models, and incorporate more detailed modeling of relevant processes such as generation of candidate interpretations.

Data availability

The code and information about stimuli access are available on Github: https://github.com/goldengua/Cognition-Noisy-channel-ERP.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cognition.2022.105359.

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.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2008). An alternative perspective on "semantic P600" effects in language comprehension. *Brain Research Reviews*, *59*(1), 55–73.

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, C., & Hagoort, P. (1993). The processing nature of the N400: Evidence from masked priming. *Journal of Cognitive Neuroscience*, *5*(1), 34–44.

Calloway, R. C., & Perfetti, C. A. (2017). Integrative and predictive processes in text reading: The N400 across a sentence boundary. *Language, Cognition and Neuroscience*, *32*(8), 1001–1016.

Cheimariou, S., Farmer, T. A., & Gordon, J. K. (2019). Lexical prediction in the aging brain: The effects of predictiveness and congruency on the N400 ERP component. *Aging, Neuropsychology, and Cognition*, *26*(5), 781–806.

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.

DeLong, K. A., & Kutas, M. (2020). Comprehending surprising sentences: sensitivity of post-N400 positivities to contextual congruity and semantic relatedness. *Language, Cognition and Neuroscience*, *35*(8), 1044–1063.

DeLong, K. A., Quante, L., & Kutas, M. (2014). Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia*, *61*, 150–162.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117–1121.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint*, URL http://arxiv.org/abs/1810.04805.

Dufour, S., Brunellière, A., & Frauenfelder, U. H. (2013). Tracking the time course of word-frequency effects in auditory word recognition with event-related potentials. *Cognitive Science*, *37*(3), 489–507.

Ehrenhofer, L., Lau, E., & Phillips, C. (2022). A possible cure for 'N400 blindness' to role reversal anomalies in sentence comprehension. (Manuscript available online).

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, *20*(5), 540–551.

Ettinger, A. (2018). *Relating lexical and syntactic processes in language: bridging research in humans and machines* (Ph.D. thesis).

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, *8*, 34–48.

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research*, *1146*, 75–84.

Frank, S., Otten, L., Galli, G., & Vigliocco, G. (2013). Word surprisal predicts N400 amplitude during reading. In *Proceedings of the 51st annual meeting of the association for computational linguistics* (pp. 878–883). Sofia, Bulgaria: Association for Computational Linguistics.

Frank, S., 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.

Frisch, S., Schlesewsky, M., Saddy, D., & Alpermann, A. (2002). The P600 as an indicator of syntactic ambiguity. *Cognition*, *85*(3), B83–B92.

Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, *44*(3), Article e12814.

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.

Gibson, E., & Pearlmutter, N. J. (2000). Distinguishing serial and parallel parsing. *Journal of Psycholinguistic Research*, *29*(2), 231–240.

Gouvea, A. C., Phillips, C., Kazanina, N., & Poeppel, D. (2010). The linguistic processes underlying the P600. *Language and Cognitive Processes*, *25*(2), 149–188.

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

Hagoort, P., Baggio, G., & Willems, R. M. (2009). Semantic unification. In *The cognitive neurosciences* (4th Ed.). (pp. 819–836). MIT Press.

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.

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.

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. *Probabilistic Linguistics*, *21*.

Kaan, E., Harris, A., Gibson, E., & Holcomb, P. (2000). The P600 as an index of syntactic integration difficulty. *Language and Cognitive Processes*, *15*(2), 159–201.

Kaan, E., & Swaab, T. Y. (2003). Repair, revision, and complexity in syntactic analysis: An electrophysiological differentiation. *Journal of Cognitive Neuroscience*, *15*(1), 98–110.

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.

Kolk, H. H., Chwilla, D. J., Van Herten, M., & Oor, P. J. (2003). Structure and limited capacity in verbal working memory: A study with event-related potentials. *Brain and Language*, *85*(1), 1–36.

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.

Kuperberg, G. R., Choi, A., Cohn, N., Paczynski, M., & Jackendoff, R. (2010). Electrophysiological correlates of complement coercion. *Journal of Cognitive Neuroscience*, *22*(12), 2685–2701.

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, *4*(12), 463–470.

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

LeBrun, B., Sordoni, A., & O'Donnell, T. J. (2021). Evaluating distributional distortion in neural language modeling. In *International conference on learning representations*.

Leckey, M., & Federmeier, K. D. (2020). The P3b and P600 (s): Positive contributions to language comprehension. *Psychophysiology*, *57*(7), Article e13351.

Levy, R. (2008). A noisy-channel model of human sentence comprehension under uncertain input. In *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 234–243).

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, *106*(50), 21086–21090.

Lewis, R. L. (2000). Falsifying serial and parallel parsing models: Empirical conundrums and an overlooked paradigm. *Journal of Psycholinguistic Research*, *29*(2), 241–248.

Lindborg, A., & Rabovsky, M. (2021). Meaning in brains and machines: Internal activation update in large-scale language model partially reflects the N400 brain potential. *43*, In *Proceedings of the annual meeting of the cognitive science society*. (43).

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*(4), 676.

McClelland, J. L. (1986). The TRACE model of speech perception. *Cognitive psychology*, *18*(1), 1–86.

McCoy, R. T., Min, J., & Linzen, T. (2020). BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In *Proceedings of the third blackboxnlp workshop on analyzing and interpreting neural networks for nlp* (pp. 217–227).

Van de Meerendonk, N., Indefrey, P., Chwilla, D. J., & Kolk, H. H. (2011). Monitoring in language perception: Electrophysiological and hemodynamic responses to spelling violations. *Neuroimage*, *54*(3), 2350–2363.

Merkx, D., & Frank, S. L. (2021). Human sentence processing: Recurrence or attention? In *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 12–22).

Michaelov, J., Bardolph, M., Coulson, S., & Bergen, B. (2021). Different kinds of cognitive plausibility: why are transformers better than RNNs at predicting N400 amplitude? *43*, In *Proceedings of the annual meeting of the cognitive science society*. (43).

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).

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.

Min, J., McCoy, R. T., Das, D., Pitler, E., & Linzen, T. (2020). Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 2339–2352).

Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., et al. (2020). Dissociable effects of prediction and integration during language comprehension: evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society B*, *375*(1791), Article 20180522.

Nieuwland, M. S., & Van Berkum, J. J. (2005). Testing the limits of the semantic illusion phenomenon: ERPs reveal temporary semantic change deafness in discourse comprehension. *Cognitive Brain Research*, *24*(3), 691–701.

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

Osterhout, L., Holcomb, P. J., & Swinney, D. A. (1994). Brain potentials elicited by garden-path sentences: evidence of the application of verb information during parsing.. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 786.

Poppels, T., & Levy, R. (2016). Structure-sensitive noise inference: Comprehenders expect exchange errors.. In *CogSci*.

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., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. Retrieved from https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

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.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3982–3992).

Ryskin, R., Futrell, R., Kiran, S., & Gibson, E. (2018). Comprehenders model the nature of noise in the environment. *Cognition*, *181*, 141–150.

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*, Article 107855.

Szewczyk, J. M., & Schriefers, H. (2018). The N400 as an index of lexical preactivation and its implications for prediction in language comprehension. *Language, Cognition and Neuroscience*, *33*(6), 665–686.

Van Berkum, J. J. (2009). The neuropragmatics of'simple'utterance comprehension: An ERP review. In *Semantics and pragmatics: From experiment to theory* (pp. 276–316). Palgrave Macmillan.

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443.

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.

Van Schijndel, M., & Linzen, T. (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, *45*(6), Article e12988.

Vissers, C. T. W., Chwilla, D. J., & Kolk, H. H. (2006). Monitoring in language perception: The effect of misspellings of words in highly constrained sentences. *Brain Research*, *1106*(1), 150–163.

Wicha, N. Y., Bates, E. A., Moreno, E. M., & Kutas, M. (2003). Potato not pope: human brain potentials to gender expectation and agreement in Spanish spoken sentences. *Neuroscience Letters* , *346*(3), 165–168.

Yano, M. (2018). Predictive processing of syntactic information: evidence from event-related brain potentials. *Language, Cognition and Neuroscience*, *33*(8), 1017–1031.

Yu, L., & Ettinger, A. (2020). Assessing phrasal representation and composition in transformers. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 4896–4907).

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