Cognitive Psychology 111 (2019) 15–52

Contents lists available at ScienceDirect

Cognitive Psychology

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

Language ERPs reflect learning through prediction error propagation

Hartmut Fitz[a,b], Franklin Chang[c,d,*]

[a] Donders Centre for Cognitive Neuroimaging, Radboud University Nijmegen, the Netherlands
[b] Neurobiology of Language Department, Max Planck Institute for Psycholinguistics Nijmegen, the Netherlands
[c] Kobe City University of Foreign Studies, Japan
[d] ESRC International Centre for Language and Communicative Development, United Kingdom

ARTICLE INFO

Keywords:
Event-related potentials
Error back-propagation
N400
P600
Semantic P600
Learning
Comprehension
Connectionist model
Development
Linguistic adaptation

ABSTRACT

Event-related potentials (ERPs) provide a window into how the brain is processing language. Here, we propose a theory that argues that ERPs such as the N400 and P600 arise as side effects of an error-based learning mechanism that explains linguistic adaptation and language learning. We instantiated this theory in a connectionist model that can simulate data from three studies on the N400 (amplitude modulation by expectancy, contextual constraint, and sentence position), five studies on the P600 (agreement, tense, word category, subcategorization and garden-path sentences), and a study on the semantic P600 in role reversal anomalies. Since ERPs are learning signals, this account explains adaptation of ERP amplitude to within-experiment frequency manipulations and the way ERP effects are shaped by word predictability in earlier sentences. Moreover, it predicts that ERPs can change over language development. The model provides an account of the sensitivity of ERPs to expectation mismatch, the relative timing of the N400 and P600, the semantic nature of the N400, the syntactic nature of the P600, and the fact that ERPs can change with experience. This approach suggests that comprehension ERPs are related to sentence production and language acquisition mechanisms.

1. Introduction

It is currently not known how neural activity in the brain implements mental processes. However, more is known about the linguistic version of this mind-body problem (Kim, 1998) due to experimental work on event-related potentials (ERPs). These are averaged, time-locked brain signals recorded through electroencephalography (EEG) that are linked to the mental operations supporting language processing (Kutas & Hillyard, 1980). ERPs have shown that the language system is sensitive to the mismatch between its expectations and linguistic input, and distinct components that differ in their timing are produced in response to different types of expectation violations. Two of the most extensively studied sentence-level ERP components are the *N400* and the *P600*. The N400 is a negative deflection of the EEG signal with a centro-parietal scalp distribution peaking around 400 ms after stimulus onset which has been linked to semantic processing (Kutas & Hillyard, 1980, 1984; see Kutas & Federmeier, 2011 for a review). The P600, on the other hand, is a positive deflection of the EEG signal with an onset around 600 ms, which is associated with syntactic processing (Hagoort, Brown, & Groothusen, 1993; Osterhout & Holcomb, 1992; see Gouvea, Phillips, Kazanina, & Poeppel, 2010 for an overview). Even though ERPs have been used to study language for more than three decades, the interpretation of these

[*] Corresponding author at: 9 Chome-1 Gakuen Higashimachi, Nishi Ward, Kobe, Hyogo Prefecture 651-2187, Japan.
E-mail address: fchang@inst.kobe-cufs.ac.jp (F. Chang).

https://doi.org/10.1016/j.cogpsych.2019.03.002
Accepted 14 March 2019

components is still controversial (Kaan, 2007; Swaab, Ledoux, Camblin, & Boudewyn, 2013).

The present paper will attempt to explain four critical features of ERPs. One feature is that ERPs reflect a mismatch of linguistic expectations, where the signal is larger when incoming linguistic material is unexpected. For example, a sentence with an anomalous ending like *I take coffee with cream and* **dog** elicits an N400 effect, which is a greater negative deflection for the unexpected final word relative to an expected ending (e.g., *cream and* **sugar**, Kutas & Federmeier, 2011). P600 effects show a larger positive deflection for various syntactic violations relative to grammatical controls. In both cases, larger amplitude ERPs occur when expectations are violated. This sensitivity to expectation mismatch is not an obvious feature for a system to have. For example, imagine a computer where the fan would make more noise not when it was working hard, but when you did something unexpected (e.g., opening a rarely used file). A second unexplained feature of ERPs is that different components have particular, relatively fixed temporal signatures. Again, this is different from computers, where the time needed for processing depends on the amount of work to carry out (e.g., the time to open a file depends on its size). Another feature is that components with different latencies are associated with semantic or syntactic processing. Since comprehension involves the integration of syntactic and semantic cues to compute the target meaning, it is not clear why these components should be separated in time. The final feature is that ERPs change with linguistic experience and it is not obvious why mismatch signals in the brain should adapt to the input. In this paper, we argue that these features can be better understood when ERPs are viewed as traces of a prediction error-based learning mechanism that supports language acquisition and adaptation.

1.1. ERPs as error propagation for learning

ERPs have been viewed as a direct reflection of sentence comprehension processes. However, it has been difficult to reconcile results in ERP studies with other comprehension studies using self-paced reading and eye-tracking (Rayner & Clifton, 2009). In ERPs, the semantic N400 precedes the syntactic P600. But in non-ERP eye-tracking studies, syntactic processing can sometimes precede semantic processing. For example, Clifton et al. (2003) found that syntactic ambiguity influenced first fixations within 258 ms of hearing a relative clause verb and this syntactic effect was not removed by earlier animacy information (contra Trueswell, Tanenhaus, & Garnsey, 1994, but consistent with Ferreira & Clifton, 1986; Just & Carpenter, 1992). These studies provided support for models of comprehension where syntactic processing comes first (e.g., garden-path model, Frazier & Rayner, 1982). Constraint-based models of comprehension, on the other hand, allow both syntactic and semantic knowledge to influence first pass parsing (MacDonald, Pearlmutter, & Seidenberg, 1994; Trueswell et al., 1994). For example, Trueswell, Tanenhaus, and Kello (1993) found a difference in reading times at the determiner following the matrix verb due to the presence or absence of a complementizer and this syntactic effect occurred between 360 and 420 ms, which is within the time window of the semantic N400. These first pass parsing effects, which are explained by early syntactic processes in both types of non-ERP theories, are difficult to explain in comprehension theories derived from ERPs where syntactic processing typically occurs later, at around 600 ms. Thus, the absolute timing of syntactic and semantic processes is a problem for unifying comprehension theories based on ERPs and non-ERP measures.

To address this mismatch, we provide an alternative account where ERPs and non-ERP comprehension phenomena are due to distinct processes. We argue that ERPs can be explained as a learning signal that is created in the sentence production system when it is adapting to linguistic input. We will show that it is possible to explain data from twelve ERP studies using a production model without adding any ERP-specific mechanisms. To understand this account, it is necessary to look at how production representations are learned. One mechanism that explains how these representations are learned is back-propagation of error within connectionist architectures like simple recurrent networks (Elman, 1990). By spreading activation forward in the network, these systems generate predictions about the next word in a sequence and compute the *error*, which is the mismatch between predicted and actual next word (Rumelhart, Hinton, & Williams, 1986). The error is then propagated backwards through the network to change the connection weights between layers in order to make future predictions more accurate. These models can learn syntactic categories and constructions, and they can explain a range of syntactic phenomena in comprehension, production, and acquisition (Chang, 2009; Christiansen & Chater, 1999; Fitz & Chang, 2017; Reali & Christiansen, 2005; St. John & McClelland, 1990). Thus, error-based learning is one way to explain how different types of linguistic representations are acquired.

The idea that ERPs might be related to prediction error was first proposed by Gehring, Goss, Coles, Meyer, and Donchin (1993). Holroyd and Coles (2002) directly instantiated the link between ERPs and error for learning within a reinforcement learning model. Later, Rabovsky and McRae (2014) showed that word-level N400s could also be explained by prediction error. What was missing in this literature is a motivation for why prediction error ERPs are generated regularly during normal sentence comprehension. Our explanation is that ERPs reflect *linguistic adaptation* processes that occur whenever input is heard (Dell & Chang, 2014). For example, structural priming is a phenomenon where a participant's tendency to use a particular structure in production is increased by their previous experience with that structure (Bock, 1986). These changes seem to be persistent (Bock & Griffin, 2000; Bock, Dell, Chang, & Onishi, 2007; Branigan & Messenger, 2016) and have been modeled in connectionist networks by leaving error-based learning ON during the processing of the prime, which leads to weight changes that bias the system towards the prime structure (Chang, Dell, & Bock, 2006). On this account, the production system generates a covert lexical prediction during comprehension, and the difference between that prediction and the heard input creates an error signal. This lexical error is propagated through the network to layers that represent syntax and changes are made to improve prediction. Thus, if production representations are adapting to the heard input through error-based learning, then production-based prediction error should be generated automatically during language comprehension.

Here, we propose that ERPs are summary signals of brain activity that index prediction error propagation during comprehension. This *Error Propagation* account explains the four features of ERPs that we mentioned earlier. Although comprehension attempts to construct a meaning representation that matches the linguistic input, this account argues that ERPs reflect the prediction error signals needed for learning, and that is why ERP amplitudes are larger when a mismatch occurs. If this error signal is automatically used for learning, then it is natural that ERP amplitude will change in response to linguistic experience, even if the linguistic input is ungrammatical ("the language acquisition process never really stops", Dell & Chang, 2014). Although non-ERP theories disagree about the timing of the use of syntactic and semantic cues in comprehension (Clifton et al., 2003; Trueswell et al., 1994), they agree that integration of these cues is ultimately needed for comprehension. In contrast, sentence production theories have argued that the independence of syntax and semantics is an important part of our ability to be productive with language (Bock & Loebell, 1990; Chang, 2002; Dell, Oppenheim, & Kittredge, 2008). The *Error Propagation* account uses production-based representations to explain ERPs, and the independence of syntax and semantics in production helps to explain the temporal difference in ERP components. Since error signals reach the lexical-semantic system before they propagate to the syntactic system, the N400 occurs earlier than the P600. Thus, the Error Propagation account can explain the P600 using the same error-based learning mechanism that generates the N400 and supports adaptation more generally.

1.2. Overview of the Error Propagation account of ERPs

Since the Error Propagation account of ERPs is complex, we will first provide a general overview of the approach. When a person hears *I take coffee with cream and*, we assume that they generate a prediction within their production system. The production system of our model has a Sequencing Layer that maps to a Lexical Layer (Fig. 1). After hearing these words, the model might predict the word *sugar* in the Lexical Layer (black circle in Fig. 1 shows the high activation for *sugar* unit, while the white circle for *dog* is not activated). The actual final word *dog* is heard and the error is the difference between target and prediction (error is shown in Fig. 1 as squares with a minus sign inside). Since *sugar* was predicted and not heard, while *dog* was heard, but not predicted, there is error for both of these units (both Lexical Layer error squares are black). After the Lexical error is generated, it is propagated to units in the Sequencing Layer to support adaptation of syntactic representations. This schematic represents the processing and learning behavior in the brain during comprehension, and we will now describe how these error signals might be related to EEG measurements. EEG

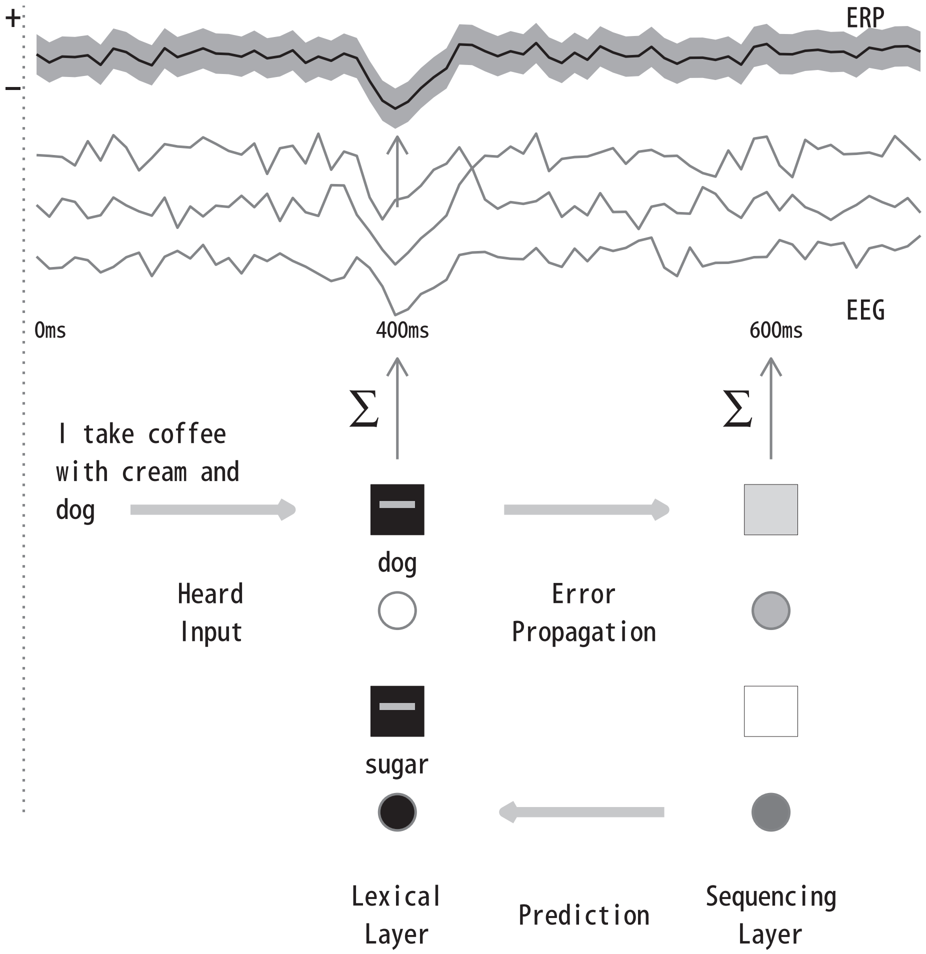

Fig. 1. Schematic of error-based learning account of language ERPs.

yields continuous value signals that summarize the activity of many synapses and neurons at particular points in time (summation symbol in Fig. 1). Our claim is that the synchronized activity during the generation and propagation of error signals at the Lexical and Sequencing Layers is reflected in the EEG signal. We also assume that it takes approximately 400 ms for the heard input to generate error at the Lexical Layer. Then it takes another 200 ms for the error to be propagated to the Sequencing Layer. If this is the case, the large error at both the *dog* and *sugar* units in the Lexical system will influence the summed EEG signal at around 400 ms. This effect will occur across multiple trials (multiple EEG waves in Fig. 1) and these are averaged to create an ERP with a negative deflection that resembles an N400. If a large amount of error is generated instead by the representations at the Sequencing Layer, then this will influence EEG at 600 ms and create a P600-like deflection in the ERP signal (this is not the case in Fig. 1, since these stimuli create an N400). This overview captures the key features of the Error Propagation account, such as the forward spread of activation to predict the next word and the backwards propagation of prediction error that generates the N400 and P600 components. In the following section we will present a more detailed example of this account to illustrate the differences between N400 and P600.

1.2.1. Error Propagation account of the N400

The N400 is a negative deflection of the EEG signal which peaks around 400 ms after the onset of a critical word. It was first found by Kutas and Hillyard (1980) in response to semantically anomalous words in sentence-final position (e.g., *He spread the warm bread with socks*). This is considered to be semantically conditioned because the N400 effect was found relative to a control sentence where the final word was semantically appropriate but in a form that was unexpected (e.g., capital letters, *She put on her high heeled SHOES*). Subsequent studies have shown that the amplitude of this component reliably indicates the degree of the semantic relationship between words and the sentence context in which they occur. For instance, the *cloze probability* of a word, defined as the proportion of subjects who complete a sentence fragment with that particular word, is the most important determinant of N400 amplitude. There is a reliable inverse correlation between cloze probability and N400 amplitude (Kutas & Hillyard, 1984) and this has been replicated many times in different languages and modalities (Besson, Faita, Czternasty, & Kutas, 1997; Federmeier, Wlotko, De Ochoa-Dewald, & Kutas, 2007; Moreno, Federmeier, & Kutas, 2002; Van Petten, Coulson, Rubin, Plante, & Parks, 1999).

Expectations about upcoming words are influenced by sentence context (Kutas & Hillyard, 1984). Context can be strongly constraining where certain words are highly likely (e.g., *The children went outside to* …), or weakly constraining in that many verb arguments are plausible (e.g., *Joy was too frightened to* …). Nevertheless, different contextual strength can create identical cloze probabilities and it has been found that N400 amplitude is determined by word expectancy rather than the strength of contextual constraints (Federmeier et al., 2007).

To see how the prediction error account can explain the N400, we trace a simplified example based on the study of Federmeier et al. (2007), where they manipulated both cloze probability (Expected, Unexpected) and contextual constraint (Strong, Weak, examples in Table 1). First we examine the hypothetical predictions in the Lexical Layer for these contexts (Predicted panel of Fig. 2). In the Strong context, a single word *play* is expected, but in the Weak context, several different words are weakly predicted (e.g., *move, play*). The expectations for these words are the same in the Unexpected condition, because the predictions are due to the preceding context, which is identical for both. The Target panel in Fig. 2 shows the target word and these words differ in the Expected (*play* and *move*) and Unexpected conditions (*look*). The Error panel in Fig. 2 shows the generated error, which is the difference between the predicted activations and the target. In the Strong Context Expected condition, the word *play* was predicted and this word was the target. The Weak Context predicts several words weakly, so there is error for the target word *move* as well as for non-target words like *play*. In contrast, in the Unexpected conditions, the target is the word *look* and there is a large negative error for the target, as well as positive error for the predicted words.

ERPs are stimulus-aligned EEG signals that reflect synaptic activity pooled from a large number of neurons in the brain (summation symbol in Fig. 1). To approximate this averaged signal, we computed the *sum* of the absolute value error (*Sum Abs. Error*) for all the units in the layer in each condition as our *linking* function for mapping the model's response onto ERPs (there is only one value for the whole layer in each condition in Sum Abs. Error panel in Fig. 2). In the Expected condition, the context directly influences the prediction error for the target words and hence there is a relatively large difference between the two contexts in the Sum Abs. Error. When the target is unexpected, its error is not modulated by the context. The Sum Abs. Error for the non-target words is similar, because there is a soft-max constraint on the Lexical system such that its activation will sum to 1. Since there is large equal amount of error for the Unexpected target and similar total absolute error for the Predicted words, this creates a large Sum Abs. Error that does not vary with context, as was found in the results of Federmeier et al. (2007).

To see this more clearly, we present model data in Fig. 3 in a way that resembles ERP waveforms where time relative to stimulus onset is shown on the x-axis. To allow us to see the two Unexpected conditions, a small amount of jitter has been added. The timing features will be explained after we introduce the model's account of the P600. For now, we can see that the Sum Abs. Error at the

Table 1
Example items from Federmeier et al. (2007).

Example sentence	Context	Expectation
The children went outside to **play**	Strong	Expected
Joy was too frightened to **move**	Weak	Expected
The children went outside to **look**	Strong	Unexpected
Joy was too frightened to **look**	Weak	Unexpected

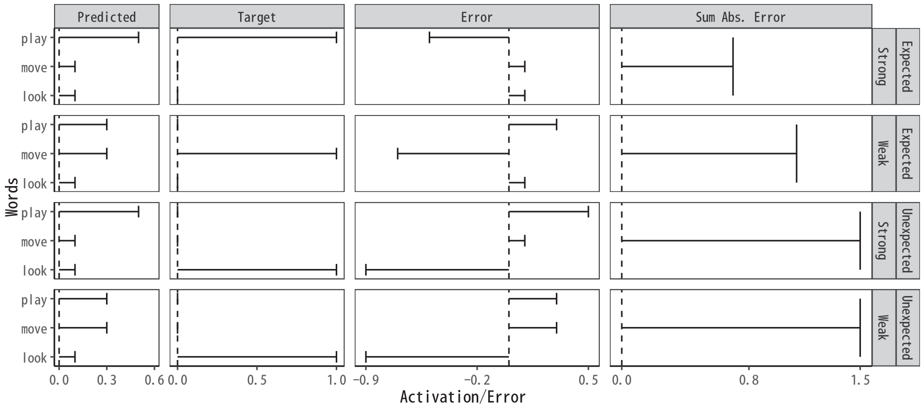

Fig. 2. Predicted/Target activation and Error/Sum Abs. Error in the Lexical Layer for Federmeier et al. (2007) study.

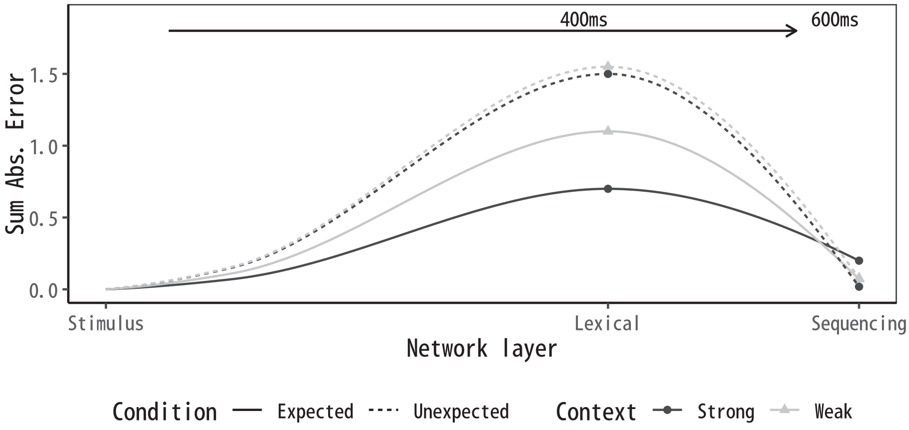

Fig. 3. ERP-type figure showing N400 pattern for Federmeier et al. (2007) study.

Lexical Layer shows a difference due to context for the Expected condition, but not the Unexpected condition, similar to the human data. This pattern is consistent with an account based on cloze probability, because cloze varies with context for expected words, but not for unexpected words. In general, however, it is not obvious why a *comprehension* effect like the N400 should be related to cloze probability, which measures the likelihood of *producing* a particular word. In the Error Propagation account proposed here, the N400 reflects error in production-based prediction, and this explains why production norms are the best predictor of mismatch signals in comprehension.

1.2.2. Error Propagation account of the P600

The other component that we will examine is the P600, which is a positive deflection of the EEG signal that peaks between 600 and 900 ms after target word onset (Gouvea et al., 2010). Relative to suitable controls, a P600 was found in response to morphosyntactic anomalies such as number agreement mismatch between the sentence subject and the main verb (Hagoort et al., 1993), violations of gender agreement (Osterhout & Mobley, 1995), tense inflection (Allen, Badecker, & Osterhout, 2003), and case marking (Coulson, King, & Kutas, 1998). Similarly, a P600 occurred in response to syntactic anomalies such as word category mismatches (Wassenaar & Hagoort, 2005), word order violations (Hagoort et al., 1993), verb subcategorization violations (Ainsworth-Darnell, Shulman, & Boland, 1998; Osterhout & Holcomb, 1992) and violations of phrase structure (Friederici, Hahne, & Mecklinger, 1996; Hagoort et al., 1993). A P600, however, was also elicited by perfectly grammatical, temporarily ambiguous garden-path sentences where the dispreferred continuation is more difficult to process than unreduced controls (Osterhout, Holcomb, & Swinney, 1994), and in unambiguous grammatical items where long-distance dependencies have to be established (Fiebach, Schlesewsky, & Friederici, 2002; Felser, Clahsen, & Münte, 2003; Kaan, Harris, Gibson, & Holcomb, 2000; Phillips, Kazanina, & Abada, 2005). Thus, the P600 indexes not only processing difficulty caused by syntactic violations, but also by temporary ambiguity and sentence complexity and its amplitude is a function of a word's syntactic fit with the preceding context.

To see how the Error Propagation account can explain the P600, we will examine the tense inflection results of Allen et al. (2003) in the model. They found a P600 when future tense auxiliaries were combined with past tense verbs *will worked*. To examine this in the model, we will compare Control sentences like *a father will* **sip** *the beer* with Violation sentences like *a father will* **sipped** *the beer*. In the model simulations, we treated morphology as a separate unit, so *sipped* is separated into its stem *sip* and the past tense morpheme *-ed*. Thus the difference between the control and violation condition is in the position after the verb *sip*. One critical feature of

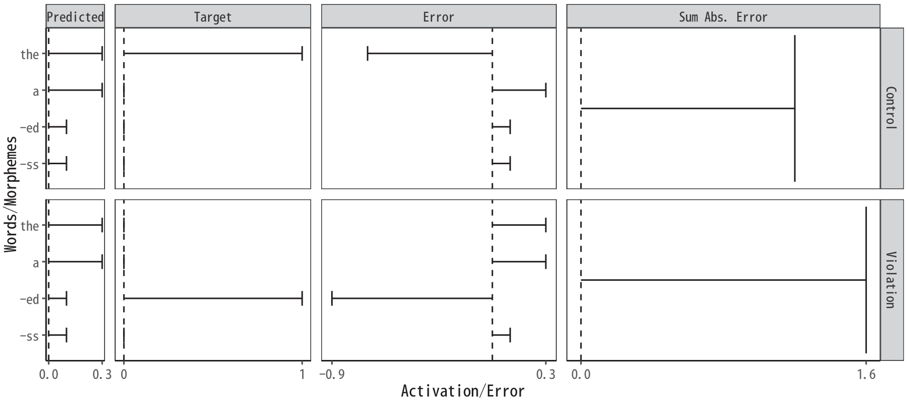

Fig. 4. Predicted/Target activation and Error/Sum Abs. Error in the Lexical Layer for Allen et al. (2003) study.

syntactic phenomena is the fact that sets of words are compatible with preceding material. For example, *will sip* can be followed by various words such as articles *a* and *the*. On the other hand, if no auxiliary was present in this context, then a third person singular morpheme *-ss* or past tense morpheme *-ed* would be acceptable continuations since the subject is singular and aspect is simple. Thus, unlike N400 phenomena where there is often only one plausible word candidate, P600 studies compare items where the context predicts categories of words. The Target panel in Fig. 4 shows that the target is *the* in the Control condition and *-ed* in the Violation condition. The Error panel shows the difference between the Predicted and Target. Since ERPs are recorded in the same way for both N400 and P600 studies, the Sum Abs. Error for the Lexical Layer is computed for both conditions.

An important feature of the error back-propagation algorithm (Rumelhart et al., 1986) is that a system could learn useful internal representations by propagating error generated at the output back to internal layers. In this process, error terms at one layer are multiplied by the connection weights projecting from the previous layer. In the present account, error at the Lexical Layer is multiplied by the weights to the Sequencing Layer and this error is used to learn the syntactic representations in the Sequencing Layer. To make this more concrete, let us assume that there are two Sequencing units S1 and S2. S1 becomes activated after verbs following future tense auxiliaries. It has positive weights of 4 to the Lexical units *the* and *a*, which can occur in this context (Fig. 5). S2 is activated after a verb when no auxiliary is present and hence it has positive weights to morphemes that can occur in this context, namely *-ed* and *-ss*. Since future and non-future tense are mutually exclusive, the model will learn negative weights between S1/S2 and the words in the other category to inhibit activation of ungrammatical continuations (these are set to −1 in Fig. 5). Normally, these weights support predictions and the arrows in Fig. 5 show the forward direction for prediction/production. But in error-based learning, error at the Lexical Layer is propagated backward in the network by multiplying the error terms by the weights and this creates the Weighted Lexical Error values for each Sequencing unit (Weighted Lexical Error panel in Fig. 6). These error values are summed for each Sequencing unit (Sequencing Error panel in Fig. 6) and this summation is an important reason for the differences between the N400 and P600. In the Control condition, S1 weakly predicted *the* and *a*, but since *the* is the target, there was negative error for the target and positive error for *a* (Error panel of Fig. 4). Both of these units are connected by a weight of 4 to the S1 unit (Fig. 5), so the Weighted Lexical Error is more negative for the target *the* and more positive for the word *a* (dark S1 lines in Weighted Lexical Error panels of Fig. 6). Since the words *-ss* and *-ed* were not predicted and not seen, the error for these words is low. The Weighted Lexical Error for the S2 links are similar to S1 links, but are smaller because the weight is only −1 (light S2 lines in Weighted Lexical Error panels of Fig. 6). In the back-propagation algorithm, the Weighted Lexical Error is summed at each Sequencing unit. Since the Weighted Lexical Errors for S1 point in different directions, these cancel when summed to create the Sequencing Error (top panel in Fig. 6). This also happens for the Weighted Lexical Errors for S2, but the values are already small.

The same process applies in the Violation condition but now there is large negative error for the unexpected *-ed* as well as positive error for the predicted words *the* and *a* (Error panel of Fig. 4). Since *-ed* has a negative weight to S1 (Fig. 5), the Weighted Lexical

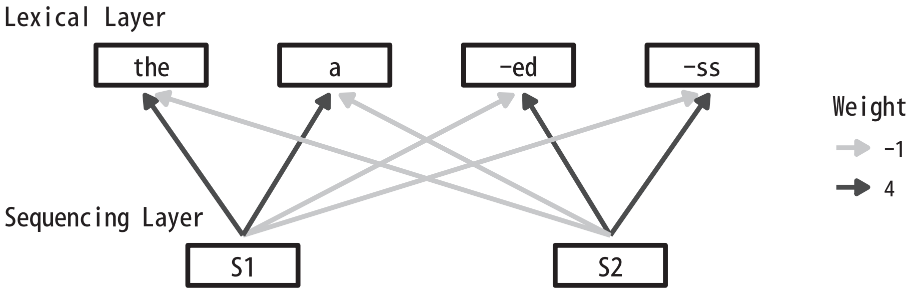

Fig. 5. Weights between Sequencing Layer and Lexical Layer. Arrows show forward direction of prediction, but error signals travel backwards and are summed at each unit in the Sequencing Layer.

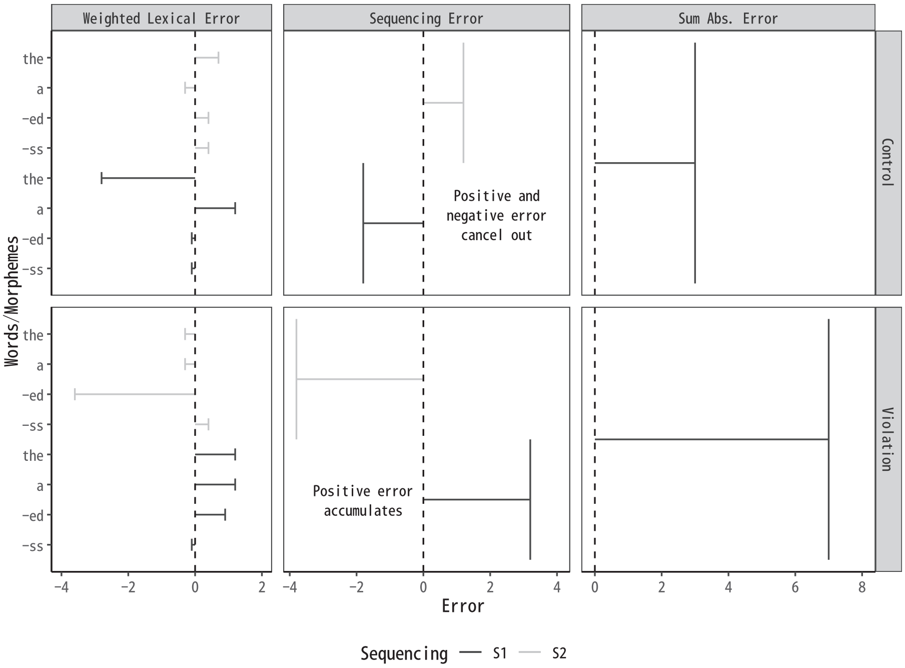

Fig. 6. Weighted Lexical Error, Sequencing Error, and Sum Abs. Error for Allen et al. (2003) study.

Errors to S1 are mostly positive. When these are summed at the Sequencing unit S1, the error will be the combined positive value. Likewise for the S2 unit, the negative error *-ed* will be multiplied by the weight of 4 to S2 (Fig. 5) and that will create a large negative Weighted Lexical Error. When error is summed at the Sequencing unit S2, it accumulates toward a large total value (bottom Sequencing Error panel in Fig. 6).

The summations at each sequencing unit are a part of error back-propagation, which is creating error signals at the Sequencing layer to support weight changes. When we apply our linking function (Sum Abs. Error), we sum error signals for all of the units within a layer to capture the fact that ERPs are summary signals from a large volume of neurons. Thus, the Sum Abs. Error panel of Fig. 6 shows a single value that combines the absolute value error for the S1 and S2 units for each condition. The Sum Abs. Error is smaller for the Control condition compared to the Violation condition, because the error cancelled in the Control condition and accumulated in the Violation condition.

We now describe how our interpolated ERP-like figures are created (Fig. 7). The Control condition has one Sum Abs. Error value at the Lexical Layer (Fig. 4) and another value at the Sequencing Layer (Fig. 6). Since we assume that it takes 400 ms to generate error at the Lexical Layer, the Sum Abs. Error value for the Lexical Layer is placed at x-position 400 ms. Likewise, the Sum Abs. Error value for the Sequencing Layer is placed at x-position 600 ms. Since ERPs are time-locked to the onset of the critical word, our ERP lines pass through 0 at time 0 ms. We also added a constraint that the value at time 100 ms would be 10% of the value of the Lexical Layer Sum Abs. Error. This helps to align model responses with patterns in human N400/P600 studies where stimulus properties require

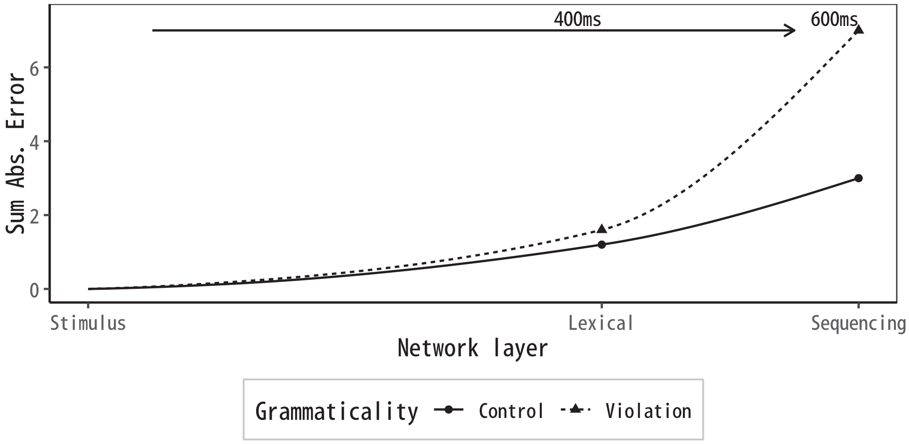

Fig. 7. ERP-type figure showing P600 pattern for Hagoort et al. (1993) study.

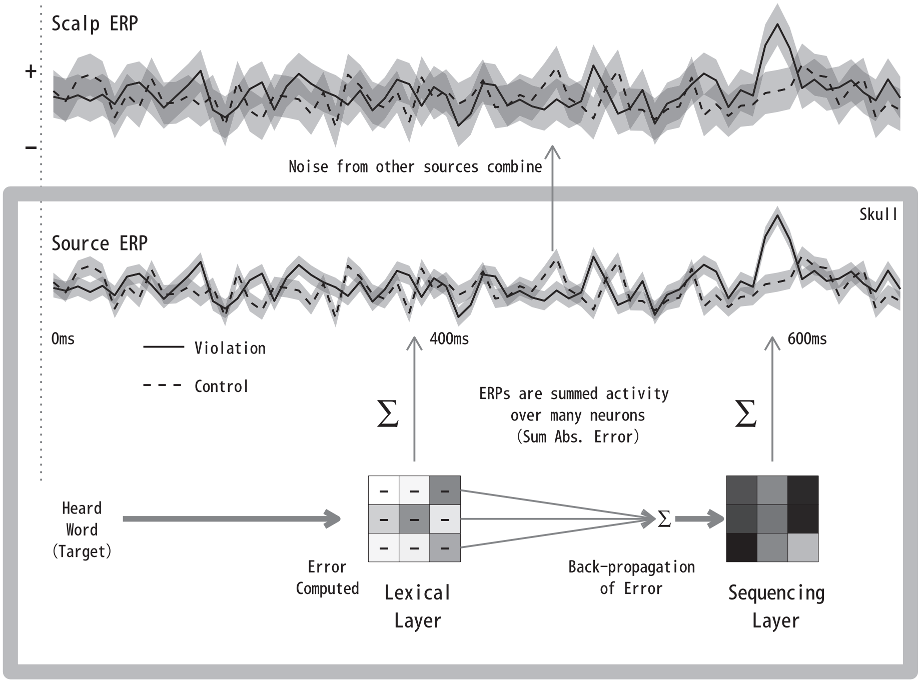

Fig. 8. Source and scalp ERPs.

some time before they begin to influence the signal (waveforms rarely diverge immediately after a word is heard). These four Sum Abs. Error values at time 0 ms, 100 ms, 400 ms, and 600 ms are fitted with piecewise cubic hermite interpolation (Fritsch & Carlson, 1980), which creates a smooth line that passes through each point without exceeding the limit values. A similar procedure was used for the Violation condition. In Fig. 7, for example, there is a larger difference in error at the Sequencing Layer compared to the Lexical Layer and this corresponds to a P600 pattern. This procedure was applied in all simulations, regardless of whether they modeled N400 or P600 effects.

1.2.3. Linking the Error Propagation account to human ERPs

Fig. 8 provides a summary of the Error Propagation account for a P600 study comparing a syntactic Violation condition with a Control condition. The Lexical Layer error is computed as the difference between the activation and the heard word target (shown by minus symbols in Fig. 8). The Sum Abs. Error is used to create the ERP signal (large summation symbol), but because the Control and Violation conditions do not strongly predict individual lexical items, there is no N400. Back-propagation of error causes the Lexical Layer error to be multiplied by the weights to the Sequencing Layer and the weighted error values are summed at each Sequencing Layer unit (small summation symbol). Sum Abs. Error is computed from the Sequencing Layer error (large summation symbol) and this creates a difference between Control and Violation in the P600 time window. The Lexical and Sequencing Layers represent brain areas that process different aspects of language, and error in these layers is akin to intracranial source ERPs. Consistent with this account, Guillem, N'Kaoua, Rougier, and Claverie (1995) examined the relationship between scalp and intracranial ERPs for the N400 and P600 in a memory task and found that different intracranial sources were associated with each component. They also showed that the skull's shielding effect caused spatial averaging of different signals, which means that ERPs recorded at the scalp contain more noise that is not related to language sources (shown by larger grey noise bands for scalp ERPs compared to source ERPs in the Fig. 8).

The difference between scalp and source ERPs has implications for how the model is compared to human data. The model's layer-specific error generates source-specific ERPs that are less noisy than human scalp ERPs and this means that we may find differences in layer-specific error, when the human scalp ERP show no difference (Source ERPs in Fig. 8 show differences between Violation and Control, that are not seen in the noisier scalp ERPs). It is non-trivial to model the additional noise in human studies as some of it is task-related (e.g., visual processing of the stimuli). Instead, we will test for an interaction of condition with the layers in the model. If there is a significant interaction, this implies that the error difference due to condition is greater in one layer than the other and therefore a noise level can be found which will make the source ERPs significantly different at one layer but not at the other. If the condition distinction is greater in the Sequencing Layer, we argue that this corresponds to a P600 component. If there is a greater effect in the Lexical Layer compared to the Sequencing Layer, we conclude that an N400 is found (the negative deflection of the source ERP is greater at 400 ms than at 600 ms in Fig. 8). This approach is used as a heuristic to be consistent across studies. In cases where human studies have found multiple components, e.g., both an N400 and P600 effect, we will conduct additional analyses to justify the claim that a biphasic pattern was also present in the model.

To summarize the account, it is assumed that listeners are constantly learning from comprehended input in order to explain syntactic adaptation behavior like structural priming. This learning mechanism is error-driven, where covert predictions are made within the sentence production system, and the error signal is used to adapt learned representations. Error differences in the Lexical Layer are assumed to be generated at around 400 ms and if a difference in conditions is found at this layer, we argue that an N400 effect is present. After another 200 ms, the error signal is propagated to the Sequencing Layer. If a large difference between conditions is found there, a P600 effect is assumed to be present. Due to the model's production architecture, the learning mechanism acquires different representations at the Lexical and Sequencing Layers and this explains why the N400 is more sensitive to lexical-semantic factors, while the P600 is more sensitive to grammatical factors. Since error is both a signal that generates ERPs and a learning signal that supports adaptation, this model naturally adapts the weights that support ERPs.

Human ERPs are a complicated phenomenon to study, because they are influenced by the individual linguistic experience of participants and the multiple electrical signals in the brain that are active during sentence processing. The goal of a computational model is to simplify this complexity to allow us to understand the core phenomena. In this section, we provided a simplified set of examples of how N400 and P600 effects are generated in our account. This allows us to trace how predictions are combined with targets to generate error which is aggregated to explain ERPs. Later we will present actual simulations demonstrating that the account can explain a range of human ERP findings within a model that makes various simplifying assumptions to allow us to understand its behavior. Before describing these simulations, however, we will contrast the Error Propagation approach with other computational models of ERPs.

1.3. Alternative accounts of ERPs

Many verbal theories have put forward a range of different mechanisms like prediction, activation, and integration to interpret ERPs (Bornkessel & Schlesewsky, 2006; DeLong, Urbach, & Kutas, 2005; Friederici, 2002; Hagoort, Baggio, & Willems, 2009; Kutas & Federmeier, 2000; Otten & Van Berkum, 2008). These mechanisms are all involved in the comprehension of meaning, which is different from the Error Propagation account that explains ERPs as learning signals. Since these comprehension mechanisms are not implemented explicitly, it can be difficult to determine how they differ from one another. However, there are several explicit computational models of ERPs which provide a clearer characterization of these mechanisms. The first set of models that we will examine are word ERP models which explain ERP processing related to single isolated words. We briefly describe these approaches in order to better understand the different ways of linking models to ERPs.

One word-level model is described by Laszlo and Plaut (2012). It was trained with back-propagation of error to map from visually presented words to semantics, and the N400 was modeled as the mean activation of the semantic output layer (see also Laszlo & Federmeier, 2011). For example, they found higher mean semantic activation for words than illegal strings and this difference was argued to be an N400 effect. Laszlo and Armstrong (2014) refined this approach to explain the effect of repetition on N400 magnitude. Cheyette and Plaut (2017) further extended the approach to cover a wider range of phenomena such as word frequency, semantic richness, priming, and orthographic neighborhood size effects. These models incorporate biologically motivated assumptions about connectivity and when they are combined with an appropriate training regime, they can show peaks around the N400 time window. Since the timing of the N400 is shaped by excitatory and inhibitory weights, it can change in response to the input that the model has been trained on.

Another approach to link models with ERPs is to use prediction error. Rabovsky and McRae (2014) developed a model of the N400 within a word-level connectionist network, where ERPs reflected prediction error between the semantics that is activated by a word and the target semantics of that word. Using an attractor network where activation would gradually approach the targets, they generated predictions based on the input word and after 10 ticks (stipulated to represent approximately 400 ms), they presented the target to measure the generated error. For example, if the word *dog* was presented to the network, it would generate a prediction for 2526 semantic feature units. Then the actual meaning of DOG (2526 semantic features derived from a norming study, McRae, Cree, Seidenberg, & McNorgan, 2005) would be activated by another part of the model to create a target signal, and the error between the model's predicted meaning and the target meaning would correspond to the N400. This model was able to explain a range of word-based effects related to frequency, repetition, number of semantic features, and neighborhood size. In three of their simulations, they left back-propagation learning ON and the model adapted its representations in response to the input. One of the main issues for this model is that it needs to guess the meaning of words in order to generate N400s, even though it already knows their target meaning.

Given the problematic assumptions with respect to the target semantics, Rabovsky, Hansen, and McClelland (2018) developed a new approach to explain sentence-level N400 effects based on the Sentence Gestalt model (St. John & McClelland, 1990). This model has two networks: Update and Query. The Update network maps from localist word units into a layer that stores the overall sentence meaning called the Sentence Gestalt. The Query network takes thematic role probe queries (e.g., AGENT) and the system tries to use the information in the Sentence Gestalt to generate 176 semantic features for the probed role. The training language did not include articles and morphology, and prepositions were cliticized to nouns (e.g., *at-breakfast man eats eggs in-kitchen*). By removing function words, the model does not learn syntactic regularities and that is why it is only sensitive to semantics. In training, each word or words was presented to the Update network and activation spread to the Sentence Gestalt. Then the Query network used the Sentence Gestalt together with the probe role to activate the lexical semantic features (e.g., location role predicts features for *kitchen*). The difference between model activation and target was used as the error signal and back-propagated through the Query network. The error for all of the role-semantic pairs in the sentence meaning was combined at the Sentence Gestalt and this combined Query network error was used to train the Update network. Through this training procedure, the Update network develops weights that attempt to store semantic information from the unfolding sentence in the Sentence Gestalt in order to support probe queries in the

Query network. To link this model to ERPs, they used the forward change in the Sentence Gestalt activation from time $t - 1$ to t, which they called the *Semantic Update* (the difference in predicted meaning before and after a word is heard). They found that Semantic Update could model various N400 effects related to semantic congruity, cloze probability, sentence position, reversal anomalies, semantic and associative priming, categorically related incongruities, and lexical frequency. They also modeled repetition priming which is a kind of linguistic adaptation. Since participants in ERP studies do not know the intended meaning of utterances before they hear them, they were not able to use the combined Query network error training procedure to model linguistic adaptation. Instead, they used a temporal difference learning approach where the Sentence Gestalt activation at the next word was used as the target for the layer at the current timestep in back-propagation. The temporal difference learning procedure would not be sufficient by itself to train the Sentence Gestalt representation to encode the input language, because the initial representation was random and training the model with random targets will not yield a semantic representation of sentence meaning. Hence this model used *different* procedures for language learning (combined Query network error) and linguistic adaptation (temporal difference error).

A related approach is offered by Brouwer, Crocker, Venhuizen, and Hoeks (2017), who provide a sentence-level model that uses back-propagation to map between localist word inputs and semantic feature representations of utterance meaning. This model has a Retrieval component that maps words and a sentence context to 100 lexical-semantic features, and an Integration component that maps lexical-semantic features onto the meaning of the whole utterance. They first trained the Integration component, then froze that system and trained the Retrieval component. The link between the model and ERP components is made in terms of the change in activations before and after a word is processed (*Activation Update*), with the N400 being the lexical semantic feature activation update and the P600 the utterance meaning activation update. They showed that the model could explain the N400 and P600 results from a single study by Hoeks, Stowe, and Doedens (2004). The Integration component is similar to the Sentence Gestalt in the Rabovsky et al. (2018) model, but the update in this layer is used to explain the P600, rather than the N400.

Another sentence-level approach that does not require target semantics is proposed by Frank, Otten, Galli, and Vigliocco (2015). They used word surprisal (the negative logarithm of conditional word probability) generated by various models to predict components in ERP waveforms. One such model was a recurrent network trained with back-propagation that predicted the next word at its output layer (Fernandez Monsalve, Frank, & Vigliocco, 2012). This layer had a soft-max activation function, so that its output would sum to 1, representing a probability distribution over words. The loss function for soft-max is proportional to the negative log activation, so this model was optimizing network weights to reduce surprisal (Jaeger & Snider, 2013). Frank et al. (2015) found that n-gram language models and recurrent networks were better at predicting N400 amplitude than models based on a probabilistic phrase structure grammar. They did not find any relationship between word or grammar-based predictions and the P600. This approach suggests that a linking theory which uses word prediction, rather than semantic feature prediction, can explain N400 patterns.

To compare these existing models with the Error Propagation account, we have highlighted their differences and commonalities in Table 2. The first distinction is whether they can explain sentence-level effects, as opposed to word-level effects. The next set of distinctions relates to their theory for linking the model to ERPs. Laszlo and colleagues (Laszlo & Plaut, 2012; Laszlo & Federmeier, 2011; Laszlo & Armstrong, 2014), as well as Cheyette and Plaut (2017), use mean or total activation at each point in time to represent N400 waveforms. The Rabovsky et al. (2018) and Brouwer et al. (2017) models use the activation change or update between two adjacent time points to encode ERPs. The Error Propagation account, as well as the models of Rabovsky and McRae (2014) and Frank et al. (2015), uses prediction error to explain ERPs. Only the Error Propagation and Brouwer et al. (2017) models account for the P600.

Another difference concerns the type of representations that are employed in linking the models to human ERPs. Most of the models postulate a Semantic Link, where activation/update/error is applied to semantic features. This account trivially explains the semantic nature of the N400, as long as appropriate semantic features are given to the model in training. The Frank et al. (2015) and Error Propagation accounts, on the other hand, use a Lexical Link, where ERPs are due to prediction error on words and hence these models must *learn* semantic word sequencing regularities. Simple recurrent networks (SRN) do learn these regularities, because syntactic and semantic categories are useful for next word prediction (see Chang, 2009, for an example of learning animacy features

Table 2
Comparison of the features of the different ERP models.

Feature	Laszlo and Federmeier (2011)	Laszlo and Plaut (2012)	Laszlo and Armstrong (2014)	Rabovsky and McRae (2014)	Frank et al. (2015)	Rabovsky et al. (2018)	Cheyette and Plaut (2017)	Brouwer et al. (2017)	Error Propagation account
Sentence Level					•	•		•	•
Activation ERPs	N4	N4	N4				N4		
Update ERPs						N4		N4/P6	
Prediction Error ERPs				N4	N4				N4/P6
Back-prop. learning	•	•	•	•	•	•	•	•	•
Semantic Link	•	•	•	•		•	•	•	
Lexical Link					•				•
Cloze Prob. N400					•				•
Adaptation = Learning				N4					N4/P6
Comprehension	•	•	•	•	•	•	•	•	

or Twomey, Chang, & Ambridge, 2014, for learning verb semantic features in the Dual-path model). For example, after the verb *eat*, all nouns might become activated to some extent, but semantically appropriate edible nouns might be activated particularly strongly (e.g. *cake*). An SRN model that predicted phonemes instead of words (e.g., Elman's (1990) letter-in-word model) would not acquire syntactic categories, because these categories cannot be signaled by the activation of multiple phoneme units which are arbitrary across categories. The existence of syntactic and semantic categories in language processing requires a mechanism for their acquisition and in SRNs, this is most easily accomplished with localist lexical targets that are the basis for the Lexical Link.

Another feature of a Lexical Link is that it is easy to explain the sensitivity of the N400 to cloze probability. Cloze probability is estimated from word production norms, but the models that use a Semantic Link are models of comprehension and do not explicitly produce words. Since there is no word production system in these models, cloze probability is just a convenient index of predictability or expectation. On the other hand, Lexical Link models predict words on the output layer and that layer has a soft-max activation function which constrains it to sum to 1, hence the predicted activation of a word unit corresponds to its cloze probability. This means that prediction error on the output layer, which is the Lexical Link to the N400, will be closely related to cloze. Semantic Link theories can model cloze effects on the N400 as probabilistic information that is distributed across many semantic features. If this approach is right, word-based probabilities like cloze should NOT predict N400 amplitude as well as semantic-feature based frequency information. This prediction has not been tested, but it would provide a way to evaluate different linking theories.

The Semantic and Lexical Link approaches are different ways to explain ERPs and they make different predictions about the processing of synonyms. Several studies have found that N400 amplitude varies for high/low cloze synonyms (e.g., Thornhill & Van Petten, 2012). Semantic Link theories explain these differences in terms of less accurate semantic feature representations for low cloze synonyms (high cloze *wheel* activates more accurate semantics features than low cloze *tire*), while Lexical Link theories explain these differences in terms of the effect of cloze on word prediction. One way to control for semantic feature accuracy is to examine translated words in fluent bilinguals. Moreno et al. (2002) presented participants with English sentences with a high cloze final word and they varied whether this word occurred in English or Spanish (*fire* or *fuego*). Since participants were highly fluent bilinguals, the ability to map the word to its semantic features should be the same across languages (no Semantic Update). However, they found a difference in the N400 for the Spanish final word in English sentences compared to the English version of the word and this suggests that the low cloze probability for Spanish words in English sentences may be contributing to ERPs. Thus, testing words with similar meaning but different cloze probabilities is one way to compare the predictions of the Semantic and Lexical Link theories.

All of the models in Table 2 used back-propagation of error to learn their internal representations (language learning), but they differ in the way that adaptation in adults is explained. The Error Propagation account learns its representations during next word prediction and linguistic adaptation is explained by applying the same procedure on utterances in ERP studies. Rabovsky and McRae (2014) also used the same procedure for both language learning and linguistic adaptation, but adaptation required that the predicted semantic features are different from the target semantic features, and it is not clear why there are two different sets of semantic features for each concept in the lexicon. Although back-propagation is used for both learning and adaptation in the Rabovsky et al. (2018) model, they used a temporal difference update procedure for linguistic adaptation that was different from the meaning probe-query training procedure for language learning. Brouwer et al. (2017) trained the Integration and Retrieval components separately, so it is not clear how this account of language acquisition would be applied to adaptation phenomena. Finally, Delaney-Busch, Morgan, Lau, and Kuperberg (2019) have used Bayesian update procedures to explain trial-by-trial adaptation in the N400 with a Lexical Link (word surprisal). Although this approach is formally similar to error-based learning with soft-max lexical outputs, it is only an account of linguistic adaptation and does not explain how word or syntactic regularities are acquired in the first place.

Further evidence that language acquisition processes should be active in adults comes from ERP studies of novel word learning. These studies present novel words in sentence contexts (e.g., *He tried to put the pieces of the broken plate back together with* **marf**, where *marf* was a synonym of glue) without explicit definitions, and they found that N400 amplitude for these novel words changes with exposure in adults (McLaughlin, Osterhout, & Kim, 2004; Borovsky, Elman, & Kutas, 2012; Borovsky, Kutas, & Elman, 2010; Mestres-Misse, Rodriguez-Fornells, & Munte, 2007). The Rabovsky and McRae (2014) model uses back-propagation for language learning and linguistic adaptation, so it could potentially explain the changes in the N400 if the system can infer the target meaning of the novel word. The Error Propagation account on the other hand attempts to predict the word form from the preceding information by linking the novel word to an existing semantic or syntactic category that is triggered by the context. Using the same algorithm that is used in language acquisition, the lexical prediction error N400 will be reduced with even a few exposures and without the word meaning. Thus, theories of ERPs can be distinguished by how much meaning is required to explain ERPs for novel words.

In most of these models, turning off the computation of summary signals of mean activation or activation update disrupts the model's ability to explain ERPs, but no other functions are affected (e.g., comprehension of meaning can still take place). In the Error Propagation account, turning off the error computation that supports ERPs also prevents the model from learning, which means that it can no longer explain adaptation effects like structural priming (Chang et al., 2006) or second language acquisition in adults (Janciauskas & Chang, 2018). In contrast, deactivating the computation of the Semantic Update in the model of Rabovsky et al. (2018) would not affect any of the comprehension results obtained with the Sentence Gestalt model in St. John and McClelland (1990). The Rabovsky and McRae (2014) model requires that the target meaning of a word appears within 400 ms, so the model is not needed to explain word comprehension. Instead, it is a model of the N400 as guessing the meaning of words that the model already knows. Thus, within the Error Propagation account, ERPs index prediction error used for non-ERP linguistic adaptation and learning, while in the other models, the summary signals that are interpreted as ERPs play no role other than to explain electrical signals on the scalp. Furthermore, although previous approaches have modeled ERPs as prediction error (Rabovsky & McRae, 2014; Frank et al., 2015), in the Error Propagation account ERPs correspond to the learning signal that is being *computed from* prediction error (see Section 2.2). This distinction is important because a learning signal has causal consequences for the language system (acquisition and adaptation) while prediction error in itself has no such consequences.

An important difference between the error-based accounts and other theories is the claim that ERPs are due to learning mechanisms that run in parallel during comprehension, but which are partially independent of comprehension. Most theories implicitly claim that ERPs reflect word- or sentence-level comprehension processes. Comparing ERPs and eye-tracking results is difficult, because different methods are typically used in each paradigm (Rayner & Clifton, 2009). ERP studies often use fixed duration presentation paradigms, while eye-tracking studies use free reading paradigms where participants can decide how long to fixate a word. However, there are studies which control for some of this variation, by using eye-tracking with fixed duration presentations and recording ERPs to the same stimuli (Dambacher & Kliegl, 2007; Sereno, Rayner, & Posner, 1998). More importantly, there are studies which co-registered ERP and eye movement results in the same participants as they read (Dimigen, Sommer, Hohlfeld, Jacobs, & Kliegl, 2011). Both of these approaches have yielded similar results, where N400 amplitude, but not timing, is sensitive to word factors (e.g. predictability), but the same lexical factors influence the timing of eye movements. The fact that the same lexical factors have different effects on timing in ERPs and eye-tracking is still not well explained in existing theories of comprehension. Most of the sentence-level models here can explain effects of predictability. In the Rabovsky et al. (2018) Sentence Gestalt model of comprehension, its Semantic Update N400 is different for high and low constraint sentences. But eye-tracking studies have found earlier effects of constraint (Ashby, Rayner, & Clifton, 2005; Rayner & Well, 1996). For example, Ehrlich and Rayner (1981) found a difference between high and low constraint before 255 ms. Sereno and Rayner (2003) argued that the effect of context appears even earlier, as it takes about 100 ms to program an eye movement. These results suggest that constraint has an effect at 150 ms in order to drive eye movement planning, but then the system waits another 250 ms before allowing the constraint to influence the N400 amplitude. Thus, the timing of eye-tracking fixations and ERPs appear to be mismatched and this is a problem for theories that argue that ERPs are a way to study the basic processes underlying comprehension.

A similar issue concerns the relationship between sentence meaning and ERPs. For example, the Brouwer et al. (2017) model argues that the P600 indexes all changes in utterance meaning in the Integration system between time $t - 1$ and t. When the model is garden-pathed, this requires a large change in meaning to recover and this creates a P600. If the updated representation is correct, then the model will be able to answer comprehension questions accurately by querying its utterance meaning. In this case, the P600 will correlate with correct responses to comprehension questions. However, there are studies that suggest that human sentence comprehension does not exhibit this correlation. Using a Good-Enough-Processing paradigm (Ferreira & Patson, 2007), Christianson, Hollingworth, Halliwell, and Ferreira (2001) found that participants who heard sentences like *While the man hunted the deer ran into the woods* tended to incorrectly answer comprehension questions like *Did the man hunt the deer?* in the affirmative, because they retained temporary interpretations they had computed during parsing. Qian, Garnsey, and Christianson (2017) conducted an EEG version of this study and found a P600 in response to the disambiguating verb (e.g., *ran*), which provides ERP evidence that participants were garden-pathed and/or attempted a repair. But when they separated the data by whether the comprehension questions were answered correctly or not, there was no difference in the P600 effect. Thus, this study is problematic for theories that argue that the P600 reflects changes in semantic meaning that support question answering.

In addition, there are several cases where experimental variables have different effects on sentence judgements and ERPs. For instance, Kaan (2002) found that subject-verb agreement errors were sensitive to the distance between subject and verb. Yet, the P600 in the same participants showed no sensitivity to subject-verb distance. Thus, distance was an important factor in judgements, but not ERPs. A similar mismatch was found for the N400, where Fischler, Bloom, Childers, Arroyo, and Perry (1984) varied the truth value (true vs false, e.g., false: *your father is a woman*) and its strength (strong vs weak, e.g., weak true: *your favorite food is steak*) in a sentence verification task. Mean response RTs were around 800 ms and there was a main effect for truth value in that people were faster for true facts than false statements, and faster for strong than for weak facts. However, they also found an interaction of these factors for the N400, where false statements elicited a larger amplitude than true ones, and this difference was smaller in the weak items. This study shows that an interaction in ERPs around 400 ms can disappear in judgements at 800 ms. Thus, even if the timing issues with existing ERP models could be resolved, it would still be necessary to postulate two sets of representations to explain the way that distance, truth value, and strength differentially influence ERP and non-ERP measures.

To summarize, most of the above models of ERPs claim that ERPs reflect comprehension processing. But factors that influence amplitude in ERP studies appear to also influence timing in non-ERP studies. Eye-tracking exhibits earlier effects of semantic and syntactic processing than ERPs. Furthermore, ERP and non-ERP effects within the same participants do not always correlate or yield the same pattern of main effects and interactions. These mismatches are problematic for theories that assume that ERPs directly track comprehension processes. To explain these mismatches, we argue that ERPs are the result of *production-based learning processes* that are distinct from those that engaged in the comprehension of meaning.

Most ERP models view the N400 and P600 as signals related to the comprehension of meaning, and hence they use a Semantic Link to connect with human ERPs. Since ERPs in these models are related to comprehension, separate mechanisms are needed to explain changes due to learning. Thus, Rabovsky et al. (2018) uses one procedure to train the model in language acquisition (e.g., combined Query network error backpropagation), another procedure for linguistic adaptation (e.g., temporal difference learning), and a third procedure to generate N400s (e.g., semantic update). In contrast, the Error Propagation account treats all ERPs as learning signals. Since humans are not always given semantic features as inputs, learning accounts use the words that are heard to generate ERP signatures such as the prediction error on words (Lexical Link). Since linguistic adaptation takes place whenever input is comprehended and is due to the same learning mechanism underlying language acquisition, the Error Propagation account generates prediction error ERPs as a side effect of these learning processes.

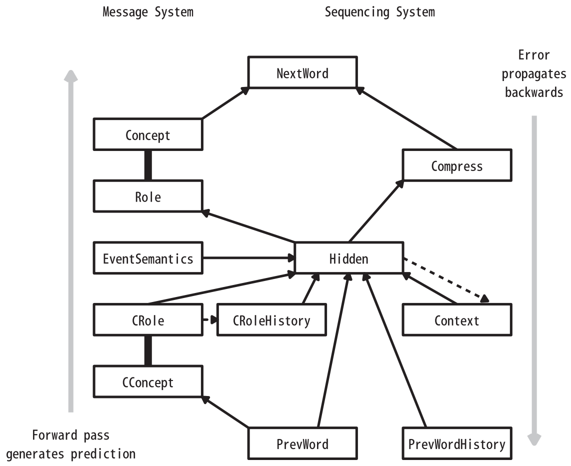

Fig. 9. The Dual-path model architecture. Word input arrives at the PrevWord layer, predicted words are activated at the NextWord layer. Solid arrows indicate connection weights that are adapted during learning, copy connections are represented by dashed arrows. Thick solid lines between roles and concepts encode the message.

2. A connectionist model of event-related potentials

To develop an explicit account of ERPs, we extended a model of language acquisition and sentence production called the Dual-path model (Chang, 2002) which has several features that make it appropriate for this task. The model has two pathways, one for sequencing and one for meaning (Fig. 9). In language acquisition, the model learns syntax and semantic regularities and associates them with different layers (Chang, 2009; Chang et al., 2006; Fitz & Chang, 2017; Twomey et al., 2014) and this can help to explain the existence of syntactic and semantic components. In addition, the model produces sentences incrementally. Thus, syntactic and semantic activations change at different points in a sentence (Fitz, Chang, & Christiansen, 2011; Chang, 2009) and this is important for explaining how ERPs vary in response to the sentence context. Furthermore, the model has been shown to be able to model linguistic adaptation phenomena like structural priming using its error-based learning mechanism (Chang et al., 2006). This requires that it generates predictions by spreading activation in the forward direction (upward arrow in Fig. 9) from the heard input in PrevWord layer. Prediction error is generated at the NextWord layer and this error is propagated backwards in the network to modify syntactic representations (downward arrow in Fig. 9). Thus, error related to semantic and syntactic representations is being generated during comprehension and this error could be used to model ERPs. These features suggest that the model might be a suitable starting point for developing an account of multiple, distinct ERP components.

The sequencing system in the Dual-path model is a simple recurrent network (Elman, 1990) that learns syntactic representations by predicting the next word in sentences, one word at a time. The difference between the model's predictions and the actual next word (error) is computed and this error is propagated through the network to adjust the weights. This learning algorithm gradually makes future predictions more accurate (Rumelhart et al., 1986). The sequencing system maps from the previous word (PrevWord layer) to a Hidden layer, that maps to the present word (NextWord layer) through a Compress layer that forces the model to develop categories. The NextWord layer corresponds to the Lexical Layer and the Hidden layer corresponds to the Sequencing Layer in the simplified model of Section 1.2. The Hidden layer receives activation from a Context layer that holds a copy of the previous Hidden layer activation state and this memory allows the model to learn sequence regularities. In the present work, we are interested in capturing cloze probabilities, which are dependent on particular word associations. Therefore, the PrevWord layer was directly connected to the Hidden layer, rather than passing through a compression layer as in previous versions of the Dual-path model. To further enhance the encoding of these sentence context regularities, we also added a PrevWordHistory layer that held a running average of the PrevWord layer activation state.

The meaning system encodes the sentence message (the meaning that the speaker is trying to convey) in fast-changing links between a Role layer and a Concept layer. For example, the message for the sentence *the boy chased the girl* would be instantiated as a fixed connection between a Role layer unit for AGENT and a Concept layer unit for BOY. Likewise there would be a fixed connection between the PATIENT Role layer unit and the GIRL Concept layer unit. Prior to production these links are set by message planning and are not changed by learning like the other links in the model. The Concept layer was linked to the NextWord layer, and this allowed the model to learn which words were associated with which concepts. The Hidden layer in the sequencing system was connected to the Role layer. Thus, the model could learn to activate the units in the Role layer which activated the message-specific concepts at particular points in a sentence. A second component of the message was a reverse message in the CConcept and CRole layers, where each link in the message was created in reverse (e.g. to match the above message, CConcept unit for BOY would be linked to CRole for AGENT). This message was set at the same time as the production message and the CRole layer received input from the PrevWord layer. This system allowed the model to determine the thematic role of the previous word and this information was propagated to the Hidden layer which supported the ability of the model to produce syntactic alternations (e.g., active-passive). To enhance memory for the roles that have already been produced, there was a CRoleHistory layer which held a running average of the CRole activations. The

Table 3
Example message-sentence pairs from model's input language.

Type	Example Message-Sentence Pair
Unaccusative	0A = BOUNCE 0Y = TOY,THE,PL 0E = PAST,PROG,AA,YY
Intransitive	*the toy -s were bounce -ing.*
Unergative	0A = WALK 0Y = GRANDMA,THE,PL 0E = PAST,SIMP,AA,YY
Intransitive	*the grandma -s walk -ed.*
Unergative	0A = JUMP 0Y = SISTER,THE 1Y = HUSBAND,A,NEAR 0E = PRES,SIMP,AA,YY
Locative	*the sister jump -ss near a husband.*
Active	0A = SNIFF 0X = TEACHER,A 0Y = WINE,THE,PL 0E = FUTURE,SIMP,AA,XX,YY
Transitive	*teacher will sniff the wine -s.*
Passive	0A = PUSH 0X = GIRL,THE 0Y = BREAD,PRN 0E = PRES,SIMP,AA,XX,**YY**
Transitive	*it is push -par by the girl.*
Believe	0A = BELIEVE 0X = FRIEND,PRN 0Y = FATHER,THE 0E = PRES,SIMP,AA,XX,YY
Transitive	*she believe -ss the father.*
Believe	0A = BELIEVE 0X = MAN,THE,PL 0E = PRES,SIMP,AA,XX,YY
Sentence	1A = WALK 1Y = DRIVER,PRN 1E = PAST,SIMP
Complement	*the man -s believe that he walk -ed.*
Prepositional	0A = SEND 0X = FATHER,A 0Y = BEER,THE,PL 0Z = BROTHER,PRN,PL
Dative	0E = PRES,SIMP,AA,XX,ZZ,YY
	a father send -ss the beer -s to them.
Double	0A = SEND 0X = BROTHER,THE 0Y = COFFEE,A 0Z = BOY,PRN
Object	0E = FUTURE,SIMP,AA,XX,YY,**ZZ**
Dative	*the brother will send him a coffee.*

Morphemes: plural *-s*, past *-ed*, third person singular *-ss*, progressive *-ing*, past participle *-par*

final part of the message was the EventSemantics layer, which held information about the number and relative prominence of arguments in the message, as well as tense and aspect information. This information was provided to the Hidden layer and helped the model to select appropriate structures in production.

This complex architecture is motivated by the need to explain a range of different behaviors from production (structural priming, Chang et al., 2006; heavy NP shift, word order effects, Chang, 2009; aphasia, Chang, 2002) and acquisition studies (auxiliary inversion, Fitz & Chang, 2017; acquisition of verb classes, Twomey et al., 2014; accessibility hierarchy, Fitz et al., 2011). It has also been applied to learn syntactic constraints from different languages (German, Chang, Baumann, Pappert, & Fitz, 2015; Japanese, Chang, 2009; English-learning Korean speakers, Janciauskas & Chang, 2018). In this work, we are testing whether ERP effects in comprehension can arise out of an architecture that was designed originally for production and acquisition.

2.1. Language input and model performance

The goal of building an explicit computational model of ERPs is to clarify the complex relationship between the brain and language processing. One of the challenges in understanding human ERPs is the large amount of variation due to the linguistic input. To reduce this variation, we used a simplified version of the English language that covers the linguistic distinctions needed to test several ERP phenomena. The model's input language was created by a symbolic grammar that generated message-sentence pairs (see Appendix B for details). Table 3 shows the different constructions in the grammar and each construction was associated with a set of abstract roles (e.g., action 0A, agents 0X, theme/patients 0Y, goals 0Z) and event-semantics information (0E).

Each role contained concepts that were appropriate for the construction (e.g., agents tend to be animate) as well as information about number and definiteness of determiners (e.g., 0X = BOY,THE,PL would generate *the boy -s*, 0Z = BROTHER,PRN,PL would generate the pronoun *them*). Event-semantics contained tense/aspect information (e.g., PRESent/PAST/FUTURE tense, SIMPle/PROGressive aspect) and information about the number of arguments (e.g., AA,XX,YY,ZZ implies that there are three arguments in addition to the action). Prominence was signaled by varying the activation of these argument markers (shown in bold in Table 3) and these were associated with structural choices (XX,YY → active transitive, XX,**YY** → passive). Transitives occurred in passive voice 20% of the time (active otherwise) and prepositional and double object forms of the dative construction occurred equally often. Believe-type verbs could occur in transitive and sentence complement forms and *that* omission occurred 25% of the time in complements. Morphemes were treated as separate words (e.g., plural *-s*). Verbs could be in past (*-ed*), present (third person singular *-ss*), or future tense as well as simple or progressive aspect (e.g., *is jump -ing*). There was also a past participle morpheme *-par*. There were other important features of the language which will be presented in more detail in the sections below.

The symbolic grammar was used to generate message-sentence pairs and the model had to learn the language from these pairs. Ten training sets were created to simulate the variability in the input across different participants, resulting in ten model subjects. Each training set contained 50,000 randomly generated sentences that were paired with messages that encoded their meaning. The model was trained for 2 sweeps through the training set for a total of 100,000 input sentences. Since meaning cannot always be inferred from the visual environment during language acquisition, a randomly selected 70% of the training items had the message removed. At the end of training, the model was tested on 200 novel sentences randomly drawn from the language. It was able to produce the exact target sentence 88% percent of the time (Bock (1986) reports human picture description accuracy at producing target structures varied between 78% and 88%).

2.2. Generating error signatures

After the model was trained on the input language, it was tested on control and violation items that mirrored the stimuli in human EEG studies. Each sentence was processed in a word-by-word fashion, activation was spread through each layer and the model attempted to predict the next word without the message. Therefore, the typical output would be a set of words corresponding to the most likely continuations of the sentence. Then each model would compute error derivatives for each layer in the same way as during learning. Most units in the model had logistic activation functions that restricted their values to between 0 and 1, but the NextWord layer used a *soft-max* activation function which created a winner-take-all bias for word selection. Since there was only one target word at each time step, the error derivative for the NextWord layer shown in (1) was simply the difference between predicted output activation y and target t for each unit j (calculation was depicted in Fig. 2; see Appendix A for details of derivation).

$$\delta_j = y_j - t_j \quad y_j,\ t_j \in \{0,\ 1\} \tag{1}$$

This error was propagated backwards in the network to generate error derivatives at deeper layers. Derivatives for these layers (e.g., Compress and Hidden) are characterized by Eq. (2) where k indexes the units in these layers and j references the units in the layer that is sending error backwards.

$$\delta_k = y_k(1 - y_k) \sum_{j=1}^{n} \delta_j w_{kj}. \tag{2}$$

The calculation in (2) was depicted earlier in Fig. 6, where the weighted Lexical error corresponds to the $\delta_j w_{kj}$ term, which is summed for each Hidden unit (the small summation symbol in Fig. 8) to create the Sequencing Error values δ_k (the $y_k(1 - y_k)$ term modulates this back-propagated error to emphasize changes where the activation is close to 0.5). This means that the P600 is sensitive to the weights between words and syntactic representations, while the N400 is more dependent on pure word expectancy. The dependent measure in our analyses was the sum of the absolute value error at each layer (the big summation symbol in Fig. 8).

For each ERP simulation, linear mixed models were fit to the data (the *bobyqa* optimizer was used and 50000 evaluations were performed; Bates, Mächler, Bolker, & Walker, 2015) and effects were sum-coded or centered (unless otherwise specified). Maximal random effect structures were fitted to each model (Barr, Levy, Scheepers, & Tily, 2013). For each of the following simulations, we crossed layer (NextWord/Hidden) against one or more factors related to the conditions in each study. If there was an interaction of layer and condition, then that provided evidence that prediction error was differentially sensitive to condition across these levels. If that interaction was significant, posthoc tests were performed for the critical contrast at the NextWord and Hidden layer separately (Hothorn, Bretz, & Westfall, 2008; Lenth, 2017). In the next sections, we will report statistical tests for simulations of several N400 and P600 studies. All of these tests involved 30 matched test items that corresponded to the linguistic manipulations in these studies. The same ten model subjects were tested across all of the following simulations (see Appendix A for details about the simulations).

2.3. Cloze probability and N400 amplitude

The most important determiner of N400 amplitude is the cloze probability of words (Kutas & Hillyard, 1984), defined as the proportion of subjects who complete a sentence fragment with that particular word. The more a word is expected in a sentence context, the smaller the N400 amplitude it generates ("congruity effect"). To simulate this relation, the model's input language contained biases for certain arguments after particular verbs. For example, the word *water* occurred 60% of the time with the verb *drink* (*High Cloze*), 15% of the time with the verb *taste* (*Medium Cloze*), and 4% of the time with the verb *take* (*Low Cloze*). Then we examined whether the model exhibited error profiles that showed a sensitivity to this knowledge by testing items such as those in Table 4.

Error was collected at the post-verbal noun position (Fig. 10). The maximal model for the data had model subject random slopes for cloze condition crossed with layer. There was a main effect of cloze, $\beta = 0.38$, SE = 0.07, $\chi^2(1) = 8.11$, p = 0.0044; and a main effect of layer, $\beta = 0.8$, SE = 0.097, $\chi^2(1) = 33.5$, p<0.001. There was also an interaction of cloze and layer, $\beta = 0.25$, SE = 0.12, $\chi^2(1) = 4$, p = 0.046, which shows that the effect of cloze was mainly centered in the NextWord layer, and in our framework this corresponds to the N400 time-window. Thus, the model acquired cloze probabilistic information about words in context and this information was encoded at the NextWord layer, not the Hidden layer.

In humans, the strength of the correlation between target word cloze and the apex of the N400 is very strong (DeLong et al., 2005 report a correlation of -0.79). In the model, NextWord layer output sums to 1.0, so the activation of each unit represents the cloze probability of the corresponding word in the context. Since the negative correlation in humans is due to the negative direction of the ERP deflection, we computed a correlation between the model's cloze probability and the negative Sum Abs. Error. The correlation is

Table 4
Example test sentences for Cloze Probability simulation.

Example sentence	Condition
a teacher was drink -ing the **water**.	High Cloze
a teacher was taste -ing the **water**.	Medium Cloze
a teacher was take -ing the **water**.	Low Cloze

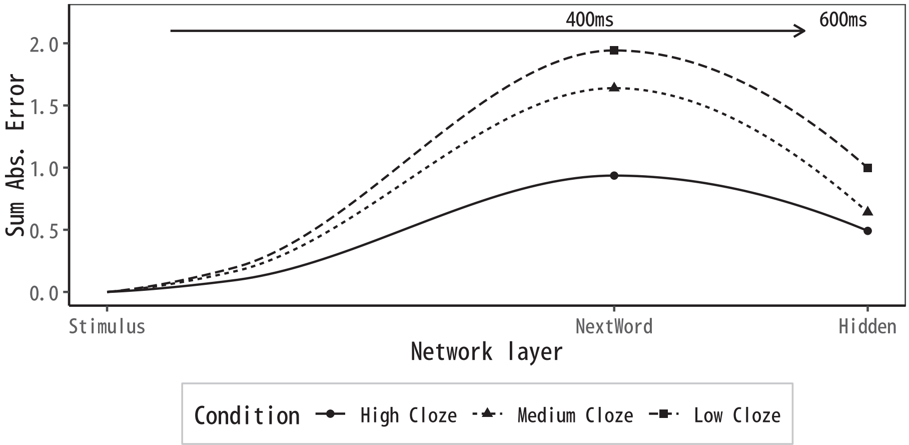

Fig. 10. Sum Abs. Error for Cloze Probability simulation.

-.999 and this tight relationship is due to Eq. (1), where cloze probability (output activation) is directly related to error (see Appendix C for more detail on this analysis). In our account, error at the NextWord layer is a source ERP and noise is added when it is recorded at the scalp. This noise can explain why the correlation between cloze and the N400 at the scalp in humans is smaller than in the model. These results show that the model replicates the inverse relationship between word expectancy and N400 amplitude in that higher cloze probabilities lead to a stronger reduction in prediction error.

2.4. *Effect of cloze probability and sentential constraint on N400 amplitude*

As discussed in Section 1.2, Federmeier et al. (2007) found that effects of sentential constraint and word expectancy (cloze probability) on the N400 were non-additive. To examine this in the model, we looked at verb-object dependencies. In the training input, the *Strong* sentential constraint verb *sip* occurred with *tea* 60% of the time and four other drinks 10% of the time. The *Weak* sentential constraint verb *sniff* occurred with *wine* 40% of the time and with the other four drinks 15% of the time. In the test items (Table 5), we varied the verbs that were used to manipulate sentential constraints for the final noun. To manipulate word expectancy, we varied whether the final noun was an *Unexpected* word like *water* or *Expected* words such as *tea* and *wine*.

Error was collected at the post-verbal noun position (Fig. 11). The maximal model for the data had all model subject random slopes for fully crossed expectancy, constraint, and layer, except for the three-way interaction and the constraint/layer interaction. There was a main effect of expectancy, $\beta = 0.45$, SE = 0.07, $\chi^2(1) = 14.78$, p<0.001; a main effect of layer, $\beta = 0.77$, SE = 0.066, $\chi^2(1) = 21.21$, p<0.001; but no effect of constraint (p = 0.602). There was also an interaction of expectancy and layer, $\beta = 0.42$, SE = 0.1, $\chi^2(1) = 10.75$, p = 0.001; an interaction of constraint and layer, $\beta = 0.27$, SE = 0.014, $\chi^2(1) = 326.97$, p<0.001; and a marginal interaction of expectation and constraint, $\beta = -0.31$, SE = 0.082, $\chi^2(1) = 3.7$, p = 0.054. Critically, there was a three-way interaction, $\beta = -0.46$, SE = 0.027, $\chi^2(1) = 268.93$, p<0.001. To break down this interaction, we performed separate models crossing constraint and expectancy at each layer. At the NextWord layer, there was a main effect of expectancy, $\beta = 0.66$, SE = 0.09, $\chi^2(1) = 17.03$, p<0.001; a marginal effect of constraint, $\beta = 0.14$, SE = 0.036, $\chi^2(1) = 3.13$, p = 0.077; and an interaction of constraint and expectancy $\beta = -0.54$, SE = 0.088, $\chi^2(1) = 16.55$, p<0.001. This interaction appears to be due to the effect of constraint in the Expected condition, but not in the Unexpected condition. At the Hidden layer, there was a main effect of expectancy $\beta = 0.24$, SE = 0.082, $\chi^2(1) = 4.54$, p = 0.033; a main effect of constraint $\beta = -0.13$, SE = 0.076, $\chi^2(1)=4.82$, p = 0.028; and no interaction (p = 0.361). This pattern of effects was similar to what has been found in the human data (Federmeier et al., 2007), where constraint did not influence the N400 for the Unexpected conditions, but did influence it for Expected conditions. They also found a late positivity in the P600 time window where the Strong Unexpected condition was different from the Weak Unexpected condition. The model showed a similar pattern, in that there was a main effect of constraint on the Hidden layer error, suggesting that the model's P600 was sensitive to constraint.

Cloze probabilities are thought to reflect semantic associations between the sentence context and a particular word. For example, the action of sipping can be done to any drink, but it is more likely for hot drinks like tea than for cold drinks like beer. In our

Table 5
Example test sentences for Cloze and Sentential Constraint simulation.

Example sentence	Constraint	Expectation
the woman will sip the **tea**.	Strong	Expected
the woman will sniff the **wine**.	Weak	Expected
the woman will sip the **water**.	Strong	Unexpected
the woman will sniff the **water**.	Weak	Unexpected

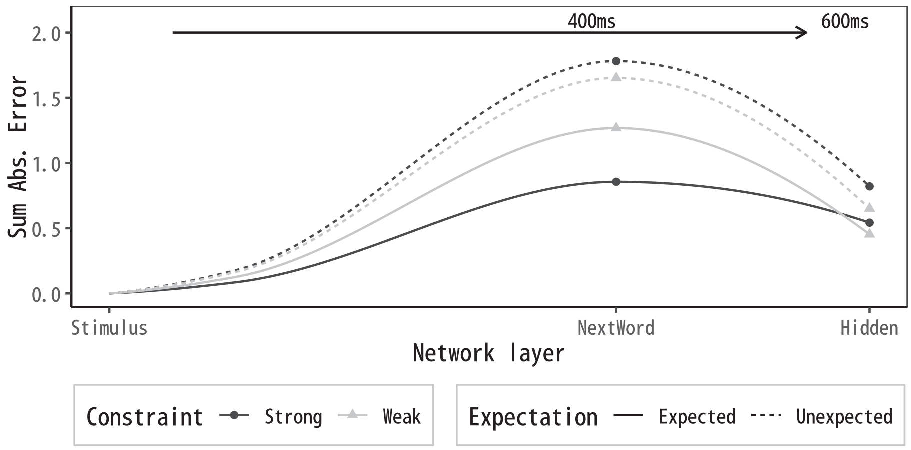

Fig. 11. Sum Abs. Error for Cloze and Sentential Constraint Simulation.

language, we created strong cloze probabilities by manipulating the pairing of verbs and arguments (e.g., *sip* and *tea* occur together 60% of the time). We also created semantic categories like drinkable liquids by having words like *tea*, *beer*, *coffee*, *water*, and *wine* occur with verbs like *sip* and *drink*. To better understand the balance between predictions for individual words and semantic categories, we took the Strong and Weak sentential constraint data for the Expected word condition and factored out how much of the Sum Abs. Error at the NextWord layer (N400) was due to the target word (e.g., *tea*) as opposed to the other members of the category (e.g., *coffee, beer, water, wine*). For the Strong constraint condition, the error was split 50% for the target word tea and the rest of the category. This was because the soft-max activation function restricted the maximum error to around 0.5. In the Weak constraint condition, there were a range of error values for both the target and the category, and there was a strong negative correlation of −0.89 between them. This means that category error was stronger when there was no clear best completion. The negative correlation demonstrates that the model was predicting semantic categories like drinks and the strength of the prediction for the category depends on whether the context was semantically biased for a single best completion.

2.5. Sentence position effects on N400 amplitude

As a sentence unfolds and a partial interpretation is being constructed, semantic information builds up and restricts options of how the sentence can continue. On a prediction-based account of the N400, word expectancy for content words should therefore increase as a function of sentence position and N400 amplitude should decrease, correspondingly. This hypothesis has been confirmed in that N400 amplitude is systematically reduced with increasing word position over the course of congruent sentences (Van Petten & Kutas, 1990; Van Petten & Kutas, 1991). In these studies, a strong inverse correlation between sentence position and N400 amplitude was observed. Words in late positions elicited smaller N400 amplitudes than earlier words since later words have more contextual support. In syntactically legal but semantically incoherent sentences, on the other hand, N400 amplitude has been found to be consistently large and does not vary depending on position (Van Petten, 1993; Van Petten & Kutas, 1991; Van Petten, Weckerly, McIsaac, & Kutas, 1997).

To test this in the model, we used dative sentences with progressively stronger positional constraints. The subject position could be filled by both animate and inanimate nouns (e.g., passive subjects). The post-verbal position could be taken by both animate and inanimate nouns, but this was influenced by verb bias: the dative verbs *give* and *lend* were biased towards animate nouns in the post-verbal slot (double object dative bias 75%), while the other two dative verbs *throw* and *send* were biased towards inanimate nouns in the post-verbal slot (prepositional dative bias 75%). The final noun position in the double object dative was typically inanimate, while in the prepositional dative, it was typically animate. To test whether constraints are more restrictive as more of the sentence is processed, we generated 30 dative test sentences (both double object and prepositional dative, which could be in active or passive voice). These *Congruent* sentences were paired with matched semantically-incoherent *Syntactic* sentences where the nouns were randomly replaced with other nouns that differed in animacy (examples are shown in Table 6). Since noun phrases could vary in terms of articles, tense, and aspect, nouns occurred in various positions in these test sentences.

Table 6
Example sentences for sentence position simulation.

Example sentence	Congruency
a **grandma** give -ed the **clerk** a **beer**	Congruent
a **pencil** give -ed the **coffee** a **friend**	Syntactic
sister -s are send -ing **toy** -s to the **father**.	Congruent
coffee -s are send -ing **man** -s to the **steak**.	Syntactic

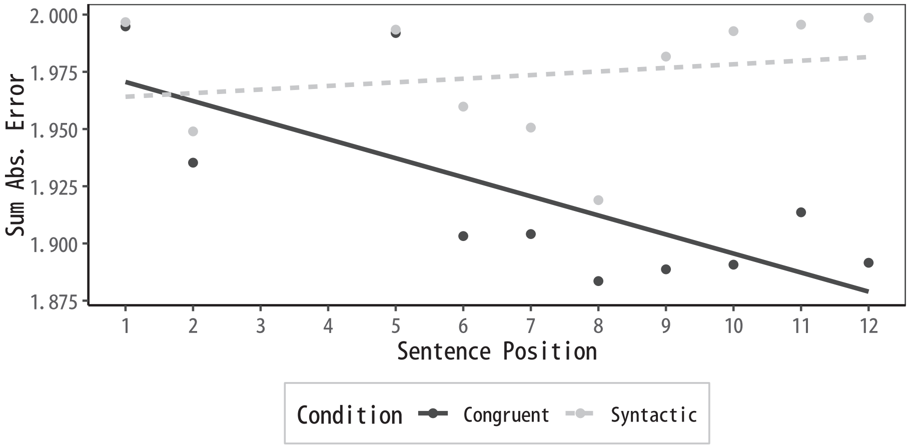

Fig. 12. Sum Abs. Error for Sentence Position simulation.

Error was collected at each noun position (Fig. 12). The maximal model for the data had random model subject slopes for congruency crossed with position. Since we were interested in the slope for the Syntactic condition, we coded congruency as a dummy coded variable (Congruent = 1) and crossed it with a centered position variable. There was a marginal positive effect of position, $\beta = 0.0027$, SE = 7e−04, $\chi^2(1) = 3.56$, p = 0.059; which suggests that there was no reduction of the slope of the Syntactic condition over sentence position (Syntactic = 0 in our treatment coding). There was a main effect of congruency, $\beta = -0.049$, SE = 0.0027, $\chi^2(1) = 8.86$, p = 0.0029, as Sum Abs. Error was higher for the Syntactic utterances. Critically, there was an interaction of congruency and position, $\beta = -0.0096$, SE = 8e−04, $\chi^2(1) = 33.16$, p<0.001, and the negative slope indicates that there was a reduction in error as sentence position increased in the Congruent condition.

Van Petten and Kutas (1991) found no effect of position in the Syntactic condition and the model is able to simulate that effect. Critically, the model showed a negative effect on error at the NextWord layer as position increased for the Congruent sentences, which mirrored the reduction in N400 amplitude over position in the human data (as in Fig. 3 in Van Petten & Kutas, 1991). The model predictions became increasingly constrained as more words were processed and this reduced the prediction error for Congruent sentences. These constraints were semantic in nature, because syntactically acceptable nouns that differed in animacy did not show the same reduction in prediction error.

2.6. *Noun-verb number agreement and the P600*

In addition to the N400, the other major component that we attempted to model is the P600, which is sensitive to a range of syntactic manipulations. One of the earliest studies to investigate ERPs in response to syntactic violations used Dutch sentences in which the subject noun mismatched the main verb in number (Hagoort et al., 1993). To simulate this study, we selected *Singular* and *Plural* third person transitive sentences with simple aspect as *Control* sentences and created *Violation* items by removing the third-person singular marker (*-ss*) for the singular subjects and added this marker for the plural subjects (Table 7).

Error was collected at the postion after the transitive verb where the sentence pairs diverged (Fig. 13). A maximal model was fit with condition and layer crossed as fixed effects and as random slopes for model subject. There was a main effect of condition, $\beta = 2.3$, SE = 0.22, $\chi^2(1) = 7.88$, p = 0.005; a main effect of layer, $\beta = 1.1$, SE = 0.31, $\chi^2(1) = 6.75$, p = 0.0094; and an interaction of condition and layer, $\beta = 3.4$, SE = 0.47, $\chi^2(1) = 19.51$, p<0.001. The interaction was due to the 6.98 times larger difference between Violation and Control in the Hidden layer, diff = 4.0225, t(9) = 9.03, p<0.001 than in the NextWord layer, diff = 0.5761, t(9) = 8.2, p<0.001. These results demonstrate that the Hidden layer was more sensitive to violations in number agreement than the NextWord layer and this instantiated a P600 pattern. This difference at the Hidden layer was the result of cancelation of positive and negative weighted lexical error in the Control condition and accumulation of error in the Violation condition when they were summed at the Hidden layer (Sequencing Layer in Fig. 6).

Table 7
Example test sentences for Noun-verb Number Agreement simulation.

Example sentence	Number	Condition
the boy take **-ss** a stick.	Singular	Control
the boy take **a** stick.	Singular	Violation
the boy -s take **a** stick.	Plural	Control
the boy -s take **-ss** a stick.	Plural	Violation

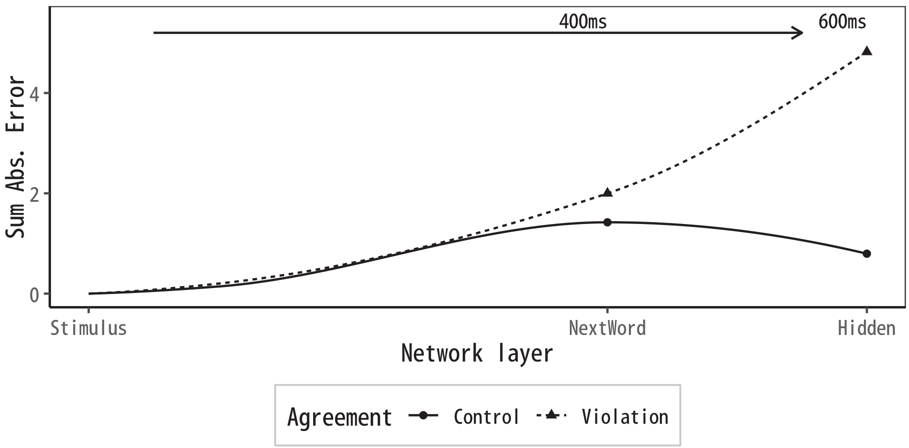

Fig. 13. Sum Abs. Error for Noun-verb Number Agreement simulation.

2.7. *Tense inflection and the P600*

In another study on inflectional morphology, Allen et al. (2003) elicited a P600 in response to violations where finite verb forms mismatched the preceding future tense auxiliaries (e.g., *The man will* **work…** versus **The man will* **worked…**). To examine this effect, we selected future tense transitive sentences with simple aspect as *Control* items and created *Violation* items by adding a past-tense marker (*-ed*) (Table 8).

Error was collected at the position after the transitive verb where the test sentences diverged (Fig. 14). The maximal model for the data had random model subject slopes for condition crossed with layer. There was a main effect of condition, $\beta = 3.5$, SE = 0.31, $\chi^2(1) = 9.92$, p = 0.0016; no main effect of layer (p = 0.746), and an interaction of condition and layer, $\beta = 5.5$, SE = 0.63, $\chi^2(1) = 22.53$, p<0.001 (Fig. 14). The interaction was due to a greater difference for condition in the Hidden layer, diff = 6.3104, t(9) = 10.19, p<0.001; than in the NextWord layer, diff = 0.7789, t(9) = 8.98, p<0.001. Since it maintained a memory of the auxiliary *will* in the Hidden layer activation, the model predicted articles after the verb and a large error was generated for *-ed*. When the error was propagated backwards, the difference at the Hidden layer was 8.1 times larger than the difference at the NextWord layer, and this created the overall P600-like pattern. As explained in Section 1.2.2, this difference was enhanced at the Hidden layer because of the cancellation and accumulation of error values during the back-propagation of error to the Hidden layer (summation in Eq. (2)).

2.8. *Word category mismatch and the P600*

Several studies have found a P600 in response to word category violations at different sentence positions (Friederici et al., 1996; Friederici, Hahne, & von Cramon, 1998; Hagoort, Wassenaar, & Brown, 2003). To simulate this in the model, our language contained some verbs such as *nap* that could also occur as nouns as in *take a nap*. We generated test sentences where these words were used as plural nouns in post-verbal position in transitives. *Violation* items were created from these *Controls* by replacing the plural marker *-s* with the past tense marker *-ed* (Table 9).

Error was collected at sentence-final position, where the test sentences differed in morphology (Fig. 15). The maximal model for the data had random model subject slopes for condition crossed with layer. There were no main effects (condition, p = 0.988, layer, p = 0.214), but there was an interaction of condition and layer, $\beta = 2.3$, SE = 0.56, $\chi^2(1) = 10.72$, p = 0.0011. The interaction was due to a significant difference in condition in the Hidden layer, diff = 2.4042, t(9) = 3.73, p = 0.0047; but no difference in the NextWord layer (p = 0.7733). One reason for the lack of a difference at the NextWord layer was the fact that the strongest prediction after the post-verbal noun was the end of sentence marker. Since this prediction generated the same error for both Controls and Violations, there was no difference at the NextWord layer. However, there was also error for past-tense *-ed*, which was magnified when back-propagated to the Hidden layer and this created the P600 in the model.

Table 8
Example test sentences for Tense Inflection simulation.

Example sentence	Condition
a father will sip **the** beer.	Control
a father will sip **-ed** the beer.	Violation

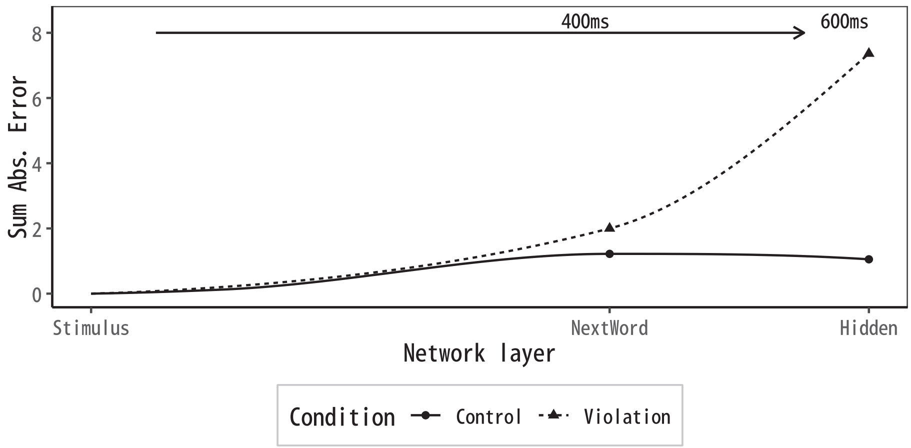

Fig. 14. Sum Abs. Error for Tense Inflection simulation.

Table 9
Example test sentences for Word Category simulation.

Example sentence	Condition
the grandma was take -ing the nap **-s**.	Control
the grandma was take -ing the nap **-ed**.	Violation

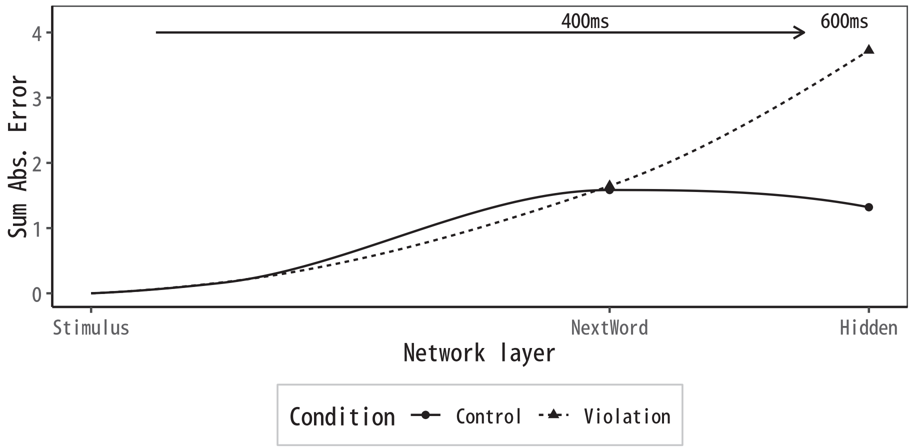

Fig. 15. Sum Abs. Error for Word Category simulation.

2.9. Verb subcategorization and the P600

Syntactic anomalies can be more subtle than a word category switch. Verb subcategorization violations, for example, are consistent with English word order and phrase structure and have been shown to elicit a robust P600 (Ainsworth-Darnell et al., 1998). In the study of Osterhout and Holcomb (1992), intransitives with a clausal complement (*The woman struggled* **to** *prepare the meal*) were compared at the infinitival marker *to* with transitives that require a direct object (**The woman persuaded* **to** *answer the door*). To examine this phenomenon in the model, we selected intransitive sentences with location adjunct phrases (*Control*) and changed the verb to a transitive one (*Violation*) which did not appear adjacent to locational adjunct phrases in the input language (Table 10).

Table 10
Example test sentences for Verb Subcategorization simulation.

Example sentence	Condition
a sister is nap -ing **near** the boy.	Control
a sister is push -ing **near** the boy.	Violation

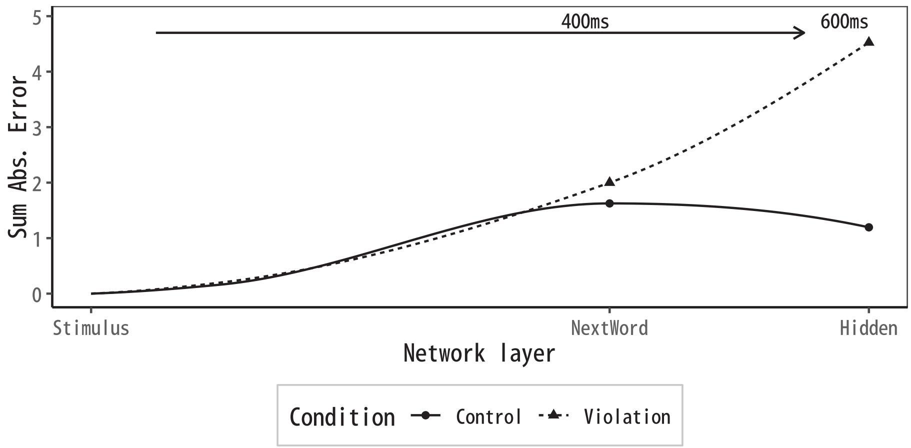

Fig. 16. Sum Abs. Error for Verb Subcategorization simulation.

Error was collected at the preposition (Fig. 16). The maximal model for the data had random model subject slopes for condition crossed with layer. There was a main effect of condition, $\beta = 1.9$, SE = 0.25, $\chi^2(1) = 7.6$, p = 0.0058; no effect of layer (p = 0.735), and an interaction of condition and layer, $\beta = 3$, SE = 0.53, $\chi^2(1) = 14.94$, p<0.001. The interaction was due to an 8.92 times larger effect of condition at the Hidden layer, diff = 3.3304, t(9) = 6.51, p<0.001; than the NextWord layer, diff = 0.3735, t(9) = 4.22, p = 0.0022. Although the model was correctly predicting prepositions after the intransitive verb at the NextWord layer, it was also strongly predicting the end of sentence marker and this created prediction error for both the Control and Violation items. Since intransitive verb representations in the Hidden layer had positive weights to both the end of sentence marker and prepositions, these positive and negative weighted error values canceled for the Control condition, creating a P600-like difference at the Hidden layer.

2.10. Garden-path sentences and the P600

The syntactic anomalies considered so far involved critical items that were strictly ungrammatical. A P600, however, has also been elicited by grammatical but dispreferred structures. Osterhout et al. (1994), for example, used garden-path sentences in which the post-verbal argument was temporarily ambiguous between a direct object and a clausal complement (*The lawyer charged the defendant was lying*) before being disambiguated by the auxiliary. Behavioral studies suggested there is a preference for the direct object interpretation (Ferreira & Henderson, 1990) which is modulated by verb biases (Garnsey, Pearlmutter, Myers, & Lotocky, 1997; Trueswell et al., 1993). The unambiguous form had an overt complementizer (*The lawyer charged that the defendant was lying*) and a P600 was found when comparing the EEG signal on the auxiliary in both items. These results showed that the P600 was not limited to outright syntactic violations but also occurred when a preferred interpretation had to be abandoned.

We created *Unambiguous* test items by generating grammatical sentences with an intransitive complement clause and the main clause verbs *believe* and *know* (Table 11). Since these verbs also occurred as transitives in the model's input language (e.g., *a father believe -ed the teacher*), removing the complementizer *that* made these sentences temporarily *Ambiguous*. These verbs occurred equally often in transitives and sentence complements, so they were equibiased.

Error was collected at the embedded verb position, where the sentence complement structure was disambiguated (Fig. 17). The maximal model for the data had random model subject slopes for condition crossed with layer. There was a main effect for condition, $\beta = 0.81$, SE = 0.17, $\chi^2(1) = 5.07$, p = 0.024; a main effect of layer, $\beta = -0.48$, SE = 0.2, $\chi^2(1) = 14.71$, p<0.001; and an interaction of condition and layer, $\beta = 1.3$, SE = 0.36, $\chi^2(1) = 8.56$, p = 0.0034. The interaction was due to a larger effect of condition at the Hidden layer, diff = 1.4331, t(9) = 4.11, p = 0.0026; than the NextWord layer, diff = 0.1809, t(9) = 4.109, p = 0.0026. The magnitude of the difference at the Hidden layer was 7.92 times bigger than the difference at the NextWord layer, and this ratio explains the significant interaction. At the NextWord layer, the model predicted a wide range of verbs and auxiliaries in the Unambiguous condition and the end of sentence marker for the Ambiguous condition, which was in part due to the high frequency of transitives in the model's language. This led to a fairly high Sum Abs. Error at the NextWord layer for both conditions with a relatively small difference between them. When back-propagated to the Hidden layer, however, there was a reduction in error in the Unambiguous condition which created the P600 pattern.

Table 11
Example test sentences for Garden-path simulation.

Example sentence	Condition
a father believe -ed that the teacher **nap** -ss.	Unambiguous
a father believe -ed the teacher **nap** -ss.	Ambiguous

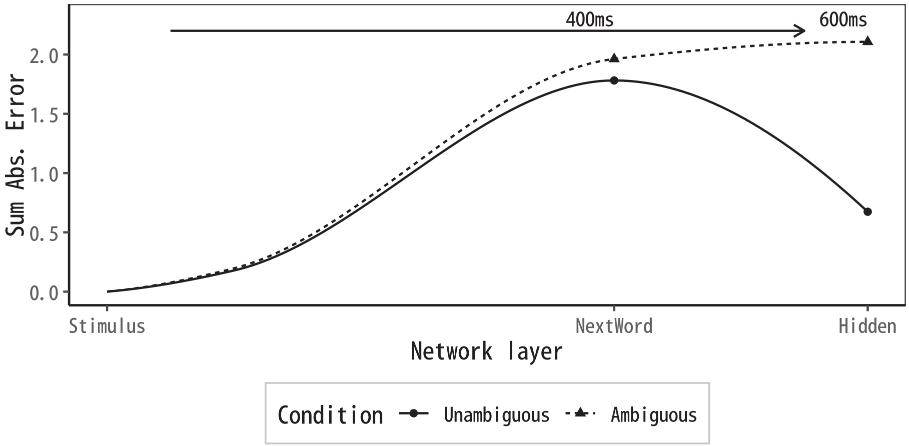

Fig. 17. Sum Abs. Error for Garden-path simulation.

2.11. The semantic P600

Several studies have investigated the processing of sentences where preferred thematic-role assignment conflicts with the argument structure of verbs (Kolk, Chwilla, van Herten, & Oor, 2003; Kuperberg, Caplan, Sitnikova, & Holcomb, 2003; Hoeks et al., 2004; Kim & Osterhout, 2005; Van Herten, Kolk, & Chwilla, 2005; Kuperberg, Kreher, Caplan, Sitnikova, & Holcomb, 2007; Nakano, Saron, & Swaab, 2010). Although the violation is semantic, these studies found a P600 on the critical words. For example, Kim and Osterhout (2005) observed a P600 effect for semantically anomalous sentences (*The hearty meal was* **devouring…**) relative to both active and passive controls (*The hungry boys were* **devouring…** and *The hearty meal was* **devoured…**). Since these items were syntactically well-formed, the effect was labelled as a *semantic* P600. To examine this effect, we randomly generated passive transitives and their active counterparts (Table 12). The role reversal condition was created from the active utterance by switching the position of the nouns.

Error was collected at the verb inflection where the sentences diverge (Fig. 18, a small amount of jitter was applied to the figure to separate the active and passive lines). The maximal model for the data had random slopes for layer for model subject. Condition was coded with a Helmert contrast where Active and Passive sentences were compared against each other, then Role Reversal was compared against the combination of Active/Passive conditions. There was a main effect of layer, $\beta = 1.3$, SE = 0.24, $\chi^2(1) = 13.7$, p<0.001. There was no difference between Active and Passive (p = 0.572), but there was a difference between the Active/Passive conditions combined and the Role Reversal condition, $\beta = 2.5$, SE = 0.026, $\chi^2(1) = 2170.95$, p<0.001. The interaction of layer and Active/Passive contrast was not significant (p = 0.567), but there was an interaction of the contrast between the Role Reversal condition and the other two conditions at the Hidden layer, $\beta = 2.5$, SE = 0.052, $\chi^2(1) = 1467.25$, p<0.001. Since the contrast between Role Reversal and the other two structures was significant overall and interacted with layer, we performed separate tests for each layer. The interaction was due to a 2.89 times larger effect of condition at the Hidden layer, diff = 5.6326, t(1787) = 102.82, p<0.001; compared to the NextWord layer, diff = 1.9456, t(1787) = 35.52, p<0.001.

In contrast to some of the previous P600 simulations, the model predicted only one morpheme in the post-verbal position, namely the progressive *-ing* morpheme in the Active Control and the past participle *-par* morpheme in the Passive Control and Role Reversal conditions. The strength of these predictions at the NextWord layer suggests that the model was using both the semantic information from the subject as well as the syntactic information in the auxiliary to make these predictions. As there was no NextWord error in the Active and Passive Control conditions to pass back to the Hidden layer, there was no difference in these conditions. The Role Reversal condition showed positive NextWord error on the *-par* morpheme, because it was wrongly predicted, while there was negative error on the *-ing* morpheme, because it was not activated enough. But since these morphemes had different syntactic distributions, they had strong weights to different Hidden layer units. Thus, when error was propagated to the Hidden layer, it did not cancel, but instead accumulated on different units and this increased the Role Reversal Sum Abs. Error at this layer associated with the P600.

The larger effect at the Hidden layer suggests that we observe a robust semantic P600 in this simulation. But in addition, the effect at the NextWord layer was 4.932 times as large as the average NextWord layer error in the five previous P600 studies. This suggests that a weaker N400 was also present in the model and this is of interest because N400s have been found in several role reversal

Table 12
Example test sentences for Semantic P600 simulation.

Example sentence	Condition
the pencil is take **-par** by the woman.	Passive Control
the woman is take **-ing** the pencil.	Active Control
the pencil is take **-ing** the woman.	Role Reversal

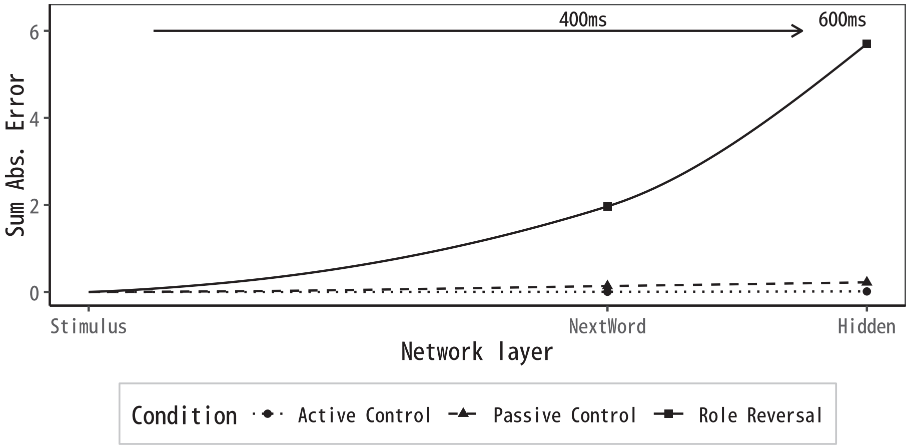

Fig. 18. Sum Abs. Error for Semantic P600 simulation.

studies (Chow & Phillips, 2013; Chow, Lau, Wang, & Phillips, 2018; Kim & Osterhout, 2005; Van Herten et al., 2005; Hoeks et al., 2004; Friederici & Frisch, 2000; Kolk et al., 2003). Kuperberg (2007) reviewed this literature and highlighted various factors that may be relevant such as semantic associations, animacy, plausibility, task, and context. The work of Chow and colleagues (Chow & Phillips, 2013; Chow et al., 2018) suggested that high predictability and distance between the arguments and the verb were also important. The model's input language was much simpler than human languages and this createed conditions of high predictability that could have enhanced the N400 effect. Future models need to be developed with more realistic inputs to better understand the relationship between the N400 and semantic P600.

2.12. Developmental changes in ERPs

Several biologically-inspired theories of ERPs have argued for a fixed association between linguistic operations and brain areas that generate ERPs (Brouwer et al., 2017; Friederici, 2002; Bornkessel & Schlesewsky, 2006). These theories predict that ERPs in children and adults should be similar and several studies support this finding (Hahne, Eckstein, & Friederici, 2004; Atchley et al., 2006), but other studies have found changes over development. Foucart and Frenck-Mestre (2012) tested a gender mismatch that generated a P600 effect in French speakers, but in English L2 learners of French, the same stimuli generated an N400. Lück, Hahne, and Clahsen (2006) found that German noun morphology errors yielded a P600 in adults and Clahsen, Lück, and Hahne (2007) found similar effects in 11–12 year olds. However, this study also found that 6–7 year olds exhibited a negativity around the N400 time window. Stronger evidence for developmental changes in the P600 comes from a study by Schneider, Abel, Ogiela, Middleton, and Maguire (2016), who presented number agreement errors to English-speaking 10–12 year old children and adults. They found a P600 and no N400 in the adults, and an N400 and no P600 in the children. While most of these studies are between-subject comparisons, Weber and Lavric (2008) examined within-subject L1 and L2 ERPs for morphosyntactic features that were similar in English (L2) and German (L1). Although participants showed a P600 in both L1 and L2, which implies that they had grammatical knowledge of both languages, they found an N400 in their L2, which was not present in their L1 or in a control group of English speakers. Overall, these studies suggest that stimuli which yield a P600 in adults can sometimes trigger an N400 in children or L2 learners. If the same linguistic distinction can activate different components over development, then that argues against a fixed association between linguistic operations and brain areas. In the Dual-path model, on the other hand, the representations that support prediction in different layers can change and reorganize over development. For example, Fitz and Chang (2017) and Twomey et al. (2014) found that the Dual-path model produced overgeneralization errors early in development and only later did the models adapt their representations to match the adult state. Since ERPs reflect the mismatch between the input and the predictions based on changing representations, the Error Propagation account predicts that it is possible for prediction-error related ERPs to also change substantially over development.

To test whether the Error Propagation account can exhibit a developmental N400, we examined the NextWord layer in the previous word category simulation (Section 2.8, Hagoort et al., 2003) at two points in the model's development (epochs 30,000 and 100,000) to see if there was a change in the N400 (Fig. 19). We did not examine the P600, since the Weber and Lavric (2008) results suggest that the developmental N400 is not inversely yoked to the development of the P600, because their highly-proficient participants showed both an N400 and a P600 in L2 English.

The maximal model for the data had random model subject slopes for condition. There was a main effect of condition, $\beta = 0.36$, SE = 0.097, $\chi^2(1) = 9.26$, p = 0.0023; a main effect of epoch, $\beta = -0.027$, SE = 0.0093, $\chi^2(1) = 6.75$, p = 0.0094; and a significant interaction of condition and epoch, $\beta = -0.3$, SE = 0.019, $\chi^2(1) = 230.02$, p<0.001. This was due to a significant difference at the child epoch 30,000, diff = 0.6558, t(9) = 6.64, p<0.001; but no difference at the adult epoch 100,000 (p = 0.5415).

These results suggest that a syntactic category violation which elicits a P600 in the adult model can also create an N400 effect in the child model. Furthermore, there were developmental changes in the N400, such that it eventually disappeared in the adult model.

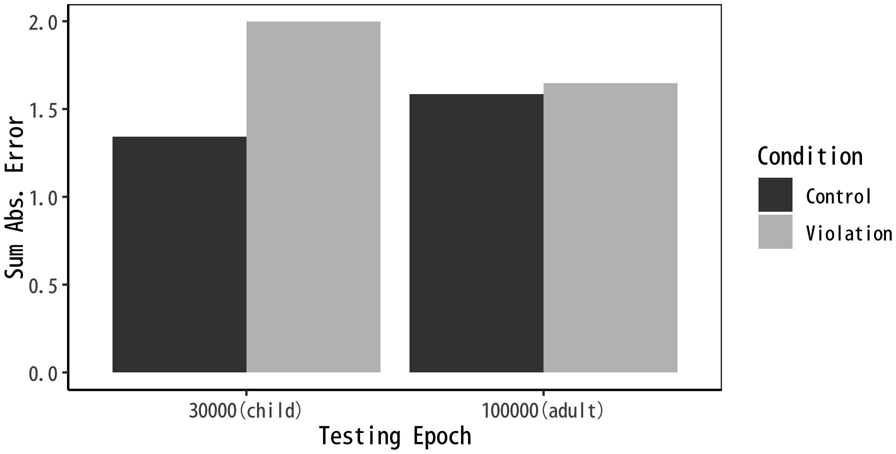

Fig. 19. Sum Abs. Error at NextWord layer for Developmental simulation.

The adult model strongly predicted the end of sentence marker after the final noun, and since that marker was not part of the Control or Violation sentence, there was no difference in error at the NextWord layer. At epoch 30,000, the prediction for the sentence final marker was reduced and the plural marker was increased and this created a difference in NextWord error (N400), since the plural marker was the final word in the Control item. This simulation exhibited a developmental N400 that eventually disappeared in the adult model. Future work is needed to understand the factors that control the appearance of the developmental N400 and how it relates to the P600 for the same stimuli.

The developmental N400 is challenging for models of the N400 which assume that the N400 can only be triggered by semantics. For example, the Rabovsky et al. (2018) model was trained with sequences without isolated function words or morphemes. Hence, this model is unable to explain why morphology can trigger a developmental N400. On the other hand, the Error Propagation model is trained with sentences that contain both function words and content words. The model has to learn lexical-semantic and syntactic constraints and assign them to particular layers in the model. Since the model does next word prediction, all syntactic knowledge is first encoded into the weights to the NextWord layer and only later are syntactic categories developed in the Compress and Hidden layers that provide a more abstract encoding of the rules of the language. Janciauskas and Chang (2018) argued that L2 learners make greater use of their lexical system for syntactic distinctions in the L2, so that might help to explain the persistent L2 N400 in bilingual adults (Weber & Lavric, 2008). Thus, the model can help to explain developmental N400 effects for syntactic stimuli in children and L2 speakers.

2.13. Linguistic adaptation of ERPs in adults

Work on linguistic adaptation has argued that language learning does not just take place during development, but also operates in response to individual linguistic experiences in adults (Dell & Chang, 2014; Dell, Reed, Adams, & Meyer, 2000; Delaney-Busch et al., 2019; Jaeger & Snider, 2013). In fact, the reason why prediction error is generated in comprehension in the present account is because of this life-long learning mechanism (Chang et al., 2006). This account predicts that the magnitude of ERPs should change in response to changes in the distribution of the items experienced in a block of trials. Evidence in support of this prediction comes from Coulson et al. (1998) who manipulated the probability of ungrammatical items in a block of trials. They tested agreement and pronoun items and varied whether there were 80% or 20% ungrammatical items in a block (the rest were grammatical). They found that the ERP amplitude around 600 ms for ungrammatical items was larger when these items were improbable (block with 20% ungrammaticality). They also found a large amplitude for grammatical items when these items were improbable (block with 80% ungrammaticality).

One way to explain this adaptation to ungrammatical items is in terms of error-based learning. To generate ERP signatures in the model, we recorded error at the NextWord layer and back-propagated it through the network. If we apply the weight updates that were computed to change the network weights, then the representations that support performance in each block will adapt to the frequency of grammatical and ungrammatical items in the block. The same approach was used in Chang et al. (2006) to explain structural priming, where learning was left ON during prime processing and that led to changes in structural representations that influenced target production. Since Coulson et al. (1998) tested subject-verb agreement items similar to those in Hagoort et al. (1993), we used our agreement items to test for adaptation in the model. Since the original study combined agreement items with pronoun items, we alternated our agreement items with an equal number of grammatical items from other control conditions as fillers (60 in total). Two matched lists of items were created where singular/plural agreement violations occurred either 80% or 20% of the time. The adult model was tested in the same way as in the other ERP simulations, but weight changes were applied after each item (the same learning rate of 0.1 as in training was used) and then these adjusted weights were used for predicting the next test item. To match Coulson et al. (1998) Fig. 5, we coded probability relative to condition (e.g., the block with 20% ungrammatical item would be

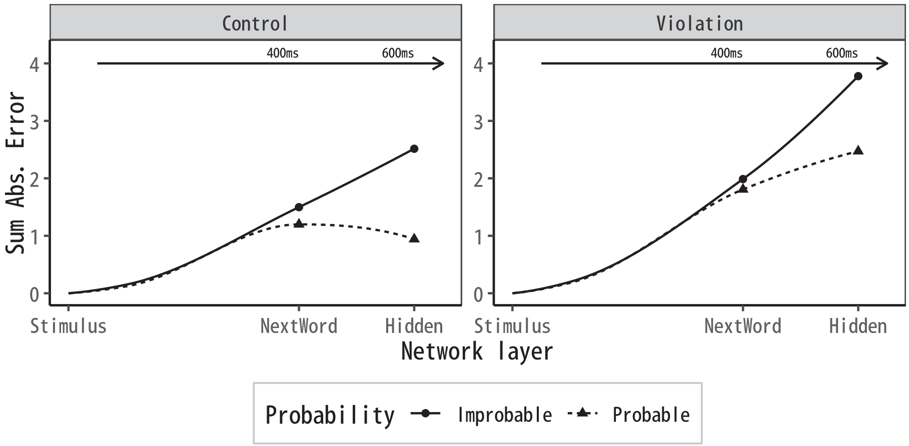

Fig. 20. Sum Abs. Error for simulation with variation in proportion of grammatical and ungrammatical stimuli.

Probable for Grammatical and Improbable for Ungrammatical) and the results are shown in Fig. 20.

We crossed condition, layer, and probability in a mixed model and the maximal model for the data had random subject slopes for condition crossed with layer and probability, except for the three-way interaction. There was a main effect of condition, $\beta = 0.97$, SE = 0.11, $\chi^2(1) = 11.25$, p<0.001; and no effect of layer (p = 0.631) or probability (p = 1.00). There was also an interaction of condition and layer, $\beta = 0.85$, SE = 0.21, $\chi^2(1) = 21.1$, p<0.001; and an interaction of layer and probability, $\beta = -1.2$, SE = 0.26, $\chi^2(1) = 12.29$, p<0.001; but no interaction of condition and probability (p = 0.752) and no three-way interaction (p = 0.471). The interaction of layer and probability was due to a 7.43 times larger effect of probability at the Hidden layer, diff = 2.4676, t(9) = 9.23, p<0.001; than at the NextWord layer, diff = 0.3323, t(9) = 1.86, p = 0.0954.

The absence of a three-way interaction or an interaction of condition and probability implies that probability had a similar effect for both Control and Violation conditions. Less probable items yielded more Sum Abs. Error at the Hidden layer and this captures the main finding of Coulson et al. (1998). The reason for the differences due to the probability of ungrammatical items is that the model was learning the likelihood of grammatical or ungrammatical transitions and that influenced the error that was propagated to the Hidden layer. An important feature of the data from Coulson et al. (1998) study was that the manipulation of probability only influenced amplitude, but not timing. In models where the timing of ERPs is shaped by learning (Laszlo & Plaut, 2012; Laszlo & Federmeier, 2011; Laszlo & Armstrong, 2014; Cheyette & Plaut, 2017), it is possible that within-experiment learning would change the timing of ERPs. On the other hand, the Error Propagation account proposes that timing is determined by the number of layers that error must be propagated across in the network, so the timing cannot be changed by experience in the same way as amplitude. This simulation shows that the error that explains ERPs can also be used to explain why ERP amplitudes change with experience.

2.14. *Priming of ERPs as error-based learning*

Repetition priming ERP studies provide evidence for the role of implicit learning in ERP adaptation (Rugg & Curran, 2007). These repetition effects can persist over lags of 120 sentences (approximately 45 min; Besson, Kutas, & Van Petten, 1992) and can appear in amnesics (Olichney et al., 2000), which suggests that they are supported by a type of implicit learning. One study that provides strong evidence for error-based adaptation is Rommers and Federmeier (2018), who examined how the repetition of the final word in paired *prime-target* sentences was influenced by the context that the final word was paired with in the earlier prime sentence. They used *target* sentences that yielded an N400 due to the fact that the final noun was only weakly predicted by the context (e.g., *It had been several years since they last cleaned the* **car**). The target was preceded by two kinds of primes with the same final word as the target (e.g., *car*). In the *Previously Predictable* prime, context predicted the final word (e.g., *Alfonso has started biking to work instead of driving his car*), while the *Previously Unpredictable* prime did not predict the final word as strongly (e.g., *Jason tried to make space for others by moving his car*). There was also an additional *Not Previously Seen* prime which had a different final word from the target (e.g., *The final score of the game was tied*). Compared to this *Not Previously Seen* prime, the primes with the same final word as the target showed a reduction in N400 magnitude. But when the prime final word was unpredictable, the N400 was reduced more than in the predictable condition and this suggests that error on the prime determined how much representations were changed. In addition to the N400, they found a late positive component around 600 ms that was also sensitive to predictability.

To simulate this study, we created *Previously Predictable* prime sentences using the high cloze probability verb-argument relationships in the language, e.g., *drink-water* (Table 13). The *Previously Unpredictable* prime sentences were the same, except that the verb was changed to the weakly constraining verb *sniff*, which occurred with *water* less frequently in the input. We also included a *Not Previously Seen* prime, which was an intransitive sentence without the word *water*. All three prime sentences were paired with the same Critical Target sentences with a weakly constraining verb *taste* and the same final noun as the *Previously Predictable/Unpredictable* primes (30 Predictable and 30 Unpredictable prime pairs). Learning was left ON during the prime and weights were updated before the Target was processed (the learning rate was 0.2). To avoid interference across prime-target pairs, weights were

Table 13
Example test sentences for Priming simulation.

Example	Condition
she drink -ss the water. .	Previously Predictable
she sniff -ss the water. .	Previously Unpredictable
a grandma will jump. .	Not Previously Seen
he is taste -ing the **water**. .	Critical Target

reset to the same adult weights before each pair as in Chang et al. (2006).

Error was collected at the final noun in the *Critical Target* item (Fig. 21). The maximal model for the data had random model subject slopes for layer. Condition was Helmert-coded to contrast *Not Previously Seen* against the other two conditions (*Seen* contrast) and another contrast that compared *Previously Predictable* against *Previously Unpredictable* (*Predictability* contrast). There was a main effect of layer, $\beta = -0.83$, SE = 0.084, $\chi^2(1) = 24.8$, $p<0.001$; a main effect of the *Seen* contrast, $\beta = 0.063$, SE = 0.0083, $\chi^2(1) = 47.66$, $p<0.001$; and a main effect of the *Predictability* contrast, $\beta = 0.14$, SE = 0.014, $\chi^2(1) = 78.44$, $p<0.001$. There was an interaction of *Seen* contrast and layer, $\beta = -0.14$, SE = 0.017, $\chi^2(1) = 67.33$, $p<0.001$; but no interaction of *Predictability* with layer ($p = 0.231$). We ran separate models with *Seen* and *Predictability* contrasts for each layer. At the NextWord layer, we found an effect of *Predictability*, $\beta = 0.15$, SE = 0.01, $\chi^2(1) = 94.25$, $p<0.001$; and an effect of *Seen*, $\beta = 0.13$, SE = 0.0058, $\chi^2(1) = 342.75$, $p<0.001$. At the Hidden layer, there was an effect of *Predictability*, $\beta = 0.12$, SE = 0.027, $\chi^2(1) = 18.85$, $p<0.001$; but no effect of *Seen* ($p = 0.684$).

In this simulation, we found that the error at the Lexical Layer was larger when the prime sentences had the same final word as the target than when they did not share this repetition. This demonstrated that repetition reduced the N400 in the model. Among the conditions with this repetition, we also found that the N400 was reduced more when the final word in the prime was unpredictable compared to when it was predictable. Since the same target sentence was used in both conditions, these effects must be due to different weights that encoded the expectations for the final word. In the model, there was larger error for the final word in the *Previously Unpredictable* prime and hence greater weight changes were made than after the *Previously Predictable* prime. These weight changes made the final word more predictable in the target and that explains why the error at the target final word was smaller in the *Previously Unpredictable* condition. Thus, the model can explain the N400 adaptation effects found in the human data. In addition, there was an effect of predictability at the Hidden layer, but no effect of repetition. In the human data, there was a late positive component in the time window of the P600 which showed a difference due to predictability in the prime, but repetition did not create a difference for both *Predictable* conditions. This simulation captures the critical link between learning and ERPs in the Error Propagation account. In the prime, next word prediction generated predictability-sensitive error signals which were used to change the linguistic representations in the system. Then, on the target, these changes influenced the prediction error that generated the N400 and that was propagated to yield a late positive component (P600).

The Error Propagation model explains adaptation as learning using a production-based error system. This error-based learning mechanism can explain production adaptation effects like structural priming (Chang et al., 2006; Jaeger & Snider, 2013; Segaert, Menenti, Weber, Petersson, & Hagoort, 2012; Tooley & Bock, 2014; Bock et al., 2007) and here we have shown that it can also explain comprehension adaptation effects in ERPs (Coulson et al., 1998; Rommers & Federmeier, 2018). Several studies have reported syntactic adaptation in non-ERP comprehension (Fine & Jaeger, 2013; Fine, Jaeger, Farmer, & Qian, 2013; Noppeney & Price, 2004; Segaert et al., 2012; Tooley & Bock, 2014; Kamide, 2012), but some of these results have not been replicated (Harrington Stack, James, & Watson, 2018; Liu, Burchill, Tanenhaus, & Jaeger, 2017). The present account suggests some reasons for this variability. In both comprehension and production tasks, we argue that speakers are generating production-based predictions about the next word

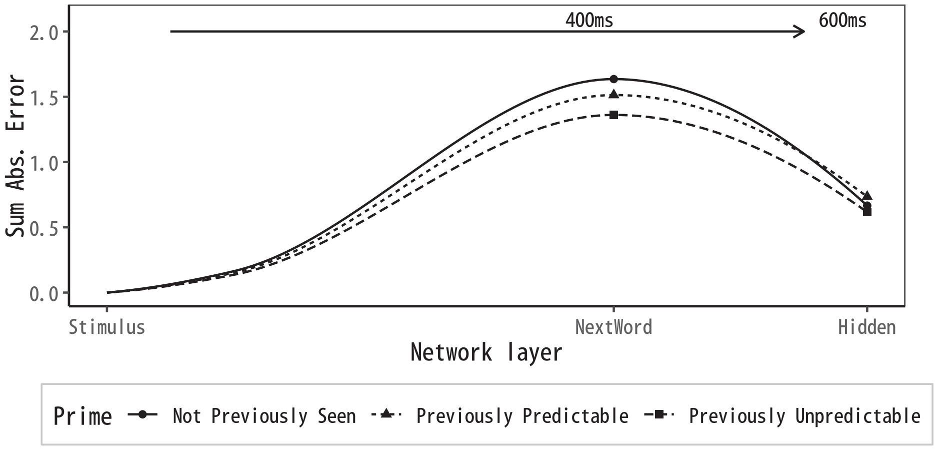

Fig. 21. Sum Abs. Error for Priming simulation.

and adjust weights that encode the structure based on prediction error. In production, the task is to generate a sequence of words and the earlier weight changes transfer directly to this task (Dell & Chang, 2014 argue that "prediction is production"). In the Error Propagation account, ERPs are caused by prediction error and syntactic adaptation is due to a learning mechanism that attempts to minimize this error. But in non-ERP comprehension tasks, the dependent measure is reading time or eye movements, and these measures are not optimized by next word prediction error. Thus, this account predicts that non-ERP comprehension adaptation will be variable.

3. Comparison of Dual-path architecture with a Simple Recurrent Network

To motivate why lexical prediction error is generated during comprehension, we used a model of sentence production (Dual-path model). In training, we occasionally included messages, which allow the model to learn to map from messages to English sentences. But in simulating ERPs, the message was turned off, since participants do not know the message until they comprehend the sentence. Therefore, when simulating ERPs, only the SRN sequencing system (Elman, 1990) was activated. To test whether the ERP results require the meaning system in the Dual-path model, we trained a separate set of models with the meaning system deactivated. These SRN models were trained with the same inputs and tested on the same items, and all of the methods and parameters described earlier were the same. The P600 results were similar in these SRN simulations, suggesting that the meaning system did not play a large role in generating these syntactic effects. The N400 results, however, were different in the SRN simulations. This is most clearly illustrated by the lack of a three-way interaction of expectation, constraint, and layer in the Federmeier et al. (2007) simulation. Even though N400s tend to elicit more error at the NextWord layer in the Dual-path model simulations, the SRN simulations showed large error at the Hidden layer in the *Unexpected Strong Constraint* condition. This is because the SRN can only predict words through its syntactic representations and hence the weights between the NextWord layer and the Hidden layer caused weighted error values to accumulate at the Hidden layer. This did not occur as much in the Dual-path model simulations, because word selection was also guided by the concepts in the message which are particularly useful for producing unexpected words. Although the message was not available during testing, its presence during training caused the Dual-path model's NextWord predictions to depend more on weights in the meaning system and hence the error for unexpected words did not have a disproportionate effect on the weights in the SRN. Gordon and Dell (2003) demonstrated that these types of models learn weights that divide the labor between syntax and meaning and it is this division of labor which explains why the meaning pathway influences word activation even when the message is not present. Although these results suggest that the Dual-path architecture may be important for explaining the N400, the fact that Elman (1990) SRN can model the P600 as a side effect of learning means that ERP predictions can be derived in the many contexts where this model has been applied.

4. Discussion

Event-related potentials have played an important role in theories that attempt to link linguistic behavior to the underlying neurobiology of the brain. Here we propose a new Error Propagation framework for understanding the N400 and P600 components which explains four crucial features of ERPs; mismatch sensitivity, semantic/syntactic dependency, their specific temporal signatures, and the adaptation of amplitude to linguistic experience. This framework assumes that the production system is constantly learning throughout one's lifetime and this involves a prediction error-based learning algorithm (Dell & Chang, 2014). Normally, the system experiences mostly grammatical inputs, which it uses to tune its representations appropriately. But occasionally, the system is placed in an ERP experiment, where it hears structures that violate its expectations and this generates large error signals which appear as ERPs. This error is propagated backwards in the network (Rumelhart et al., 1986) and since this takes time, error propagation also explains why there are components with different latencies, such as the N400 and P600. Due to the architecture of the language system, different layers encode syntax and semantics to different degrees, and the ERP components associated with these layers differ in their sensitivity to grammar and meaning. Finally, the error signal that is indexed by ERPs is used for learning and causes linguistic representations to adapt. Adaptation allows speakers/listeners to continue to use language as it changes and ERPs reflect this important function.

The Error Propagation account was able to simulate three studies on the N400: sensitivity to cloze probability (Kutas & Hillyard, 1984), the non-additive nature of sentential constraint (Federmeier et al., 2007), and amplitude reduction over sentence position (Van Petten & Kutas, 1991). The N400 was modeled using prediction error at the NextWord Layer in the Dual-path sentence production model. Since there were particular semantic associations in the model's input (e.g., people tend to sip tea), the model encoded those regularities in its NextWord Layer predictions. This layer has a *soft-max* activation function, which means that its output reflected cloze probability. Since the N400 was modeled as the error in its probabilistic predictions about next heard word, this account can explain why production cloze probability is the best predictor of N400 amplitude in comprehension.

This account was also able to simulate five studies on the P600: noun verb agreement (Hagoort et al., 1993), tense inflection (Allen et al., 2003), word category mismatch (Hagoort et al., 2003), verb subcategorization (Osterhout & Holcomb, 1992), and garden-path effects (Osterhout et al., 1994). In most of these studies, the model would predict a syntactically appropriate category of words or morphemes at the NextWord layer and this would lead to both positive and negative error terms, since only one word/morpheme could be the target. When the target was within the predicted category, the positive/negative weighted error values canceled out when summed at the Hidden layer. When the target fell outside of the predicted category, the weighted error values would change sign so that their summed value was larger at the Hidden layer. Thus, the P600 was shaped by the weights between the NextWord and Hidden layers as well as by the summation of the weighted error values at each Hidden unit.

Since error-based learning drives the model to learn the representations that allow it to best predict the upcoming input, it combined syntax and animacy information in its Hidden layer and this helped to reproduce the semantic P600 (Kim & Osterhout, 2005; Kuperberg, 2007). This accumulation of information is similar to what happens in the model's account of N400 cloze effects. The reason why error is larger at the Hidden layer is because the NextWord units that encode verb morphology are strongly connected to distinct units in the Hidden layer; some for progressive active voice, some for passive. The model learns this distinction because it is useful for future predictions (e.g., *by*-phrases tend to follow past participles). Since the Hidden units are associated with distinct syntactic choices, there will be positive weights to grammatical continuations and negative weights to ungrammatical continuations, and this leads to accumulation in the Violation condition which generates the P600.

A key issue in ERP research concerns the question why components differ in their sensitivity to syntax and semantics. Some models have the nature of their representations set by the training procedure. For example, Rabovsky et al. (2018) model was trained on sequences without separate function words or morphemes, and therefore it encoded only semantic distinctions. But if it was given input with function words and morphemes, it would encode both syntax and semantic features in its Sentence Gestalt. The Error Propagation account does not fix the model's representations through its training procedure. Instead, the Dual-path network architecture was designed to take a sequence of content and function words/morphemes and encode semantic and syntactic information into different network layers. Since the Dual-path architecture must learn its representations in acquisition, it can learn developmental representations that differ from the adult state. We found that the model initially encoded syntactic distinctions using its links to the NextWord layer and this generated a temporary N400 which disappeared later in development. This pattern matched a range of studies that have found developmental N400 effects in L1 children or L2 learners (Clahsen et al., 2007; Foucart & Frenck-Mestre, 2012; Lück et al., 2006; Schneider et al., 2016; Weber & Lavric, 2008). Thus, the same error signals that index ERP components can also help to explain how syntactic and semantic regularities are encoded separately in the first place, and why these representations might change over development.

Learning is used to both acquire syntactic and semantic representations, and also to adapt them in response to new input. As a consequence, ERPs should change as a result of linguistic experience in adults. Support for this prediction comes from linguistic adaptation studies such as Coulson et al. (1998) who found that P600 amplitude was related to the proportion of ungrammatical test items in their study. As in models of structural priming (Chang et al., 2006), it was possible to simulate these effects in the Error Propagation account by turning learning ON as the model processed the test stimuli. Error at the Hidden Layer was larger when the tested structure was improbable within the experiment and this explains adaptation in ERP amplitude in terms of the same prediction error that was used to learn internal representations. These adaptation effects could be explained by any mechanism that adapts to the input. But selective evidence in support of an error-based mechanism was provided by Rommers and Federmeier (2018), who showed that word predictability on a prime influenced how much adaptation appeared in ERPs at a later target item. There are many cases where linguistic adaptation is similar to language learning in children (e.g., novel word learning in adults Borovsky et al., 2012; Borovsky et al., 2010; McLaughlin et al., 2004; Mestres-Misse et al., 2007; Perfetti, Wlotko, & Hart, 2005), and approaches like the Error Propagation account treat these two situations as instances of the same learning procedure.

An important claim in this work is that ERPs are the result of learning processes, rather than the processes that are engaged in the comprehension of meaning. As mentioned in Section 1.3, existing theories have assumed a tight link in the mechanisms that support ERPs and non-ERP comprehension results. These theories have difficulty explaining cases where syntactic effects precede semantic effects (e.g., Clifton et al., 2003; Ferreira & Clifton, 1986; Just & Carpenter, 1992). Another issue is that the absolute timing of first pass effects of syntax typically occur before 600 ms (Frazier & Rayner, 1982; Trueswell et al., 1993, 1994). Finally, ERPs do not correlate with meaning comprehension (Qian et al., 2017), do not match subject-verb agreement errors (Kaan, 2002), and do not predict the timing of judgments based on meaning (Fischler et al., 1984). In general, ERPs do not always move in sync with non-ERP measures of comprehension.

To better understand how the Error Propagation account might explain these differences, we examine a set of studies which are closely matched in the structures tested (Garnsey et al., 1997; Osterhout et al., 1994). As discussed in the garden-path simulation (Section 2.10), Osterhout et al. (1994) found a P600 at the disambiguating embedded verb for sentence complements in ambiguous structures compared to unambiguous controls (e.g., *the judge charged (that) the defendant was lying*). The disambiguating embedded verb *was* is heard at time 0 ms (Fig. 22). The Dual-path model assumes that when the word form has been identified, it activates a unit in the PrevWord layer. Since the word that is heard at time t is the target for the prediction made at time $t-1$, the activation in the Prevword layer can be sent to the NextWord layer as a target signal and NextWord Error can be computed (NextWord activation is not shown). Then this error is back-propagated to the Hidden layer to create the Hidden Error and we saw in Section 2.10 that the larger error in the ambiguous compared to the unambiguous condition generated an P600. This sequence represents the backward pass of error which is used for learning/adaptation of the network weights and which creates ERP effects in the Error Propagation account (top pathway in Fig. 22).

To see how this account would explain eye-tracking results on the same ambiguity, we examined the findings in Garnsey et al. (1997), who found a significant difference in first pass reading times at the same disambiguating region for direct-object-biased matrix verbs (ambiguous structures 365 ms, unambiguous ones 332 ms). Hence, the same word triggers an eye movement around 300 ms, which is much earlier than the ERP occurring at 600 ms. In the Error Propagation account, the incoming word *was* is activated in the PrevWord layer and activation spreads forward to the Hidden layer (middle pathway in Fig. 22). If we assume that this forward spread of activation takes place before 600 ms, then these layers can help to explain some of the earlier syntactic effects in first pass parsing. For example, if the decision to make an eye movement is based on the information in the Hidden layer, then we might expect slower first pass times in the ambiguous condition, because the Hidden layer contains a representation for a transitive structure that conflicts with the sentence complement cue from the incoming word in the PrevWord layer. In the unambiguous

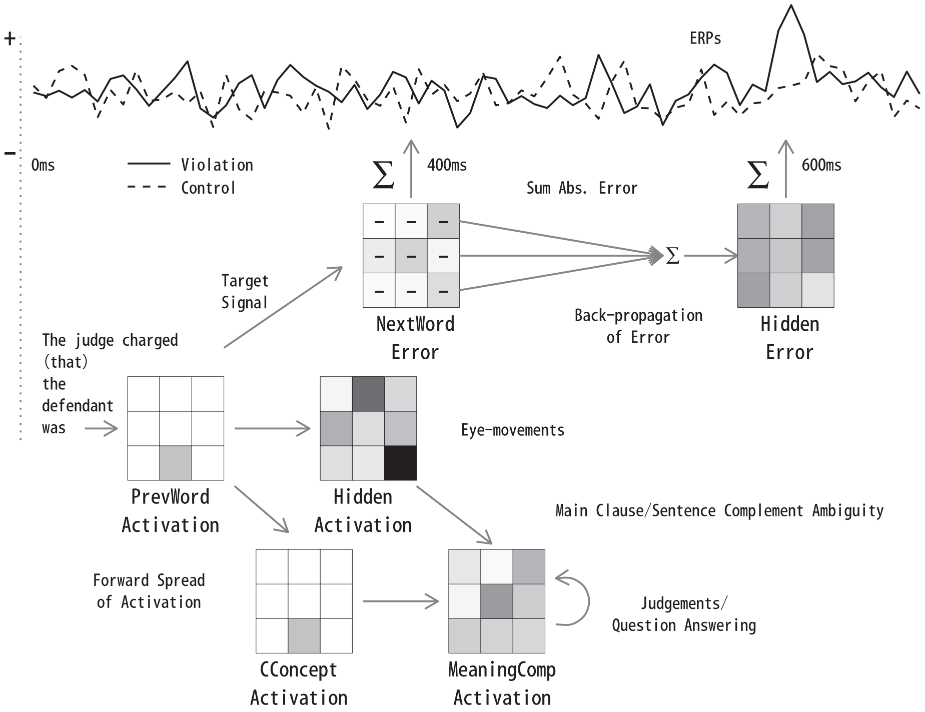

Fig. 22. Outline of extended model with meaning comprehension system.

condition, the Hidden layer already contains a representation for a sentence complement based on seeing the complementizer *that* earlier, and the structural expectations in the Hidden layer might trigger an earlier eye movement.

Another aspect of comprehension is the computation of sentence meaning and the present Dual-path model does not store or compute this kind of meaning. One way to add this ability would be to create an extra layer of MeaningComp units that would integrate syntactic material from the Hidden layer and lexical semantics from the CConcept layer (bottom pathway of Fig. 22). To allow this layer to store this information, it would need to have recurrent connections onto itself. This MeaningComp layer could support sentence judgments, and since it is separate from the NextWord Error system generating the N400, it allows for effects in judgment timings to differ from N400 effects (Fischler et al., 1984). Furthermore, as it is separate from the Hidden Error system that generates the P600, there would not need to be a strong correlation between P600 effects and correct question answering based on sentence meaning (Qian et al., 2017). More generally, the MeaningComp layer can help to explain the puzzle that is raised by Good-Enough-Processing phenomena (Ferreira & Patson, 2007), where comprehenders do not always appear to use syntactic information in computing sentence meaning (e.g., interpreting *the dog was bit by the man* as *the dog bit the man*). Why should the language system develop syntactic representations that can influence first pass parsing only to bypass them when computing sentence meaning? If the MeaningComp layer computes sentence meaning, and it gets direct inputs from the concepts in the CConcept layer, then sentence meaning can sometimes ignore syntactic cues. On this account, syntactic representations in the Hidden layer get their ultimate motivation from production, but they are used and shaped by comprehended inputs and that is why eye movements sometimes exhibit early syntactic effects (MacDonald, 2013). Thus, the Error Propagation account argues for forward and backward streams that operate on heard words in parallel with online and offline phenomena tapping different layers. While this theory is not a complete account of sentence comprehension, it can explain why there are dissociations between ERP and non-ERP studies in syntactic processing and meaning comprehension.

Compared to other models of ERPs, the Error Propagation account has several unique properties. Most ERP models only explain the N400 (Cheyette & Plaut, 2017; Frank et al., 2015; Laszlo & Federmeier, 2011; Laszlo & Plaut, 2012; Laszlo & Armstrong, 2014; Rabovsky & McRae, 2014; Rabovsky et al., 2018). Brouwer et al. (2017) model can also explain one P600 study, but the Error Propagation account explicitly captures multiple N400 and P600 effects within the same model. In contrast with the view that the P600 and N400 are independent components, the present account argues that the P600 is the result of error propagation from the system that generates the N400, so these components should be linked and this might help to explain why many studies have yielded biphasic patterns (Van Petten & Luka, 2012). Another claim of the Error Propagation account is that ERPs should adapt over trials within a study and also over development. An error-based learning account would predict, on average, that ungrammatical violation items should elicit more change than grammatical control items. While some other accounts can model adaptation (Delaney-Busch et al., 2019; Rabovsky & McRae, 2014; Rabovsky et al., 2018), it is less clear whether they could explain larger changes in ERPs such as developmental N400s, and a better understanding of the factors that generate those temporary N400s is important. Another prediction of the Error Propagation account is that production probabilities should predict N400 amplitude. Since production probabilities are group measures, it might be possible to compare production and ERPs for two groups to determine if within-group correlations are stronger than between-group correlations. The Error Propagation account also predicts dissociations between non-ERP and ERP comprehension measures. More studies using EEG and eye-tracking for the same stimuli would help to expand our understanding of the relationship between these phenomena.

Although our account provides an explanation for some of the core findings in the ERP literature, there are several limitations to this work. One limitation is that it does not explain components that occur earlier than the N400, like the N100, P200 and left anterior negativity (LAN). It is possible that some early components could be captured with prediction error by extending the model to predict phonology, but further work is needed to test this. Another limitation is that the model at present only explains a small set of N400 and P600 studies and there are many other studies that are not covered. We would like to make a distinction between phenomena which could be explained with the present set of mechanisms and those which would require different mechanisms. For example, a range of higher-order discourse factors have been found to influence ERPs (Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005; Van Berkum, Hagoort, & Brown, 1999). In sentence production, speakers must learn to use discourse information to guide lexical choice. Hence, if the message includes this information, the model's production system will generate predictions during comprehension and the resulting error will be discourse-sensitive. Thus it should be possible to exhibit discourse-related ERPs using the same basic mechanisms that we have used in this paper. On the other hand, there are ERP effects of prediction that are more difficult to explain with the present mechanism. For example, DeLong et al. (2005) presented readers with a context that generated a stronger expectation for *kite* than *airplane*. They found an N400 for the article *an* before the noun *kite* or *airplane* was actually seen (however, there are some questions about the reliability of this effect, see Nieuwland et al., 2018; Yan, Kuperberg, & Jaeger, 2017; De Long, Urbach, & Kutas, 2017). In the present model, predictions are made incrementally word-by-word, and therefore it is unlikely that the model will predict the noun two words ahead and then use that prediction to retrodict the article that precedes it. Our model is therefore just a first attempt to explain the N400 and P600 within a single account without proposing additional specialized components. It is clear that future work will be needed to develop and extend this research to provide greater coverage.

At present, many verbal theories try to explain ERPs in terms of various concepts such as prediction, activation, retrieval, repair, integration, update, and unification. What has been missing from this inventory is the idea that ERPs might reflect *learning processes* that are driven by implicit prediction error. In addition, these verbal concepts are often vague and underspecified which makes it difficult to reach consensus on the interpretation of complex, noisy, time-varying, multi-channel EEG signals for different components that change over development. Hence, another goal of this work was to provide an explicit computational theory of ERPs, which is still simple enough to understand and follow (see Section 1.2). The model activates words in the Lexical Layer during prediction from the context, retrieves the heard word, integrates/unifies it into its representation in the Sequence layer, and repairs these representations as more information is processed. All of these processes influence prediction error and help to explain the four key features of ERPs: expectation mismatch, timing, semantic/syntactic sensitivity, and adaptation to linguistic experience. On this account, ERPs are not just some summary signal to describe processing, but rather they reflect *functional* signals that the brain uses itself for learning.

The Error Propagation account argues that ERP comprehension results can be explained within a model that accounts for a broad range of other sentence production and language acquisition results in various languages. As such, it suggests a way to unify different components of the language system. In addition, the learning mechanisms that were used are domain-general (Chang, Janciauskas, & Fitz, 2012), and similar sequence-learning accounts might explain ERPs occurring outside of natural language processing (e.g., in vision, Sitnikova, Holcomb, Kiyonaga, & Kuperberg, 2008; gesture, Özyürek, Willems, Kita, & Hagoort, 2007; artificial grammar learning, Christiansen, Conway, & Onnis, 2012; music, Patel, Gibson, Ratner, Besson, & Holcomb, 1998; and mathematics, Martín-Loeches, Casado, Gonzalo, de Heras, & Fernández-Frías, 2006). In both language and non-language domains, prediction error is a useful way to learn appropriate internal representations and adaptation can explain why prediction error is generated during input processing.

When back-propagation of error was first described (Rumelhart et al., 1986), it provided a linking theory which could explain some of the mental states that generated human behavior out of the processing of simple units that were roughly modeled on the behavior of biological neurons. The optimism around this linking of the mind and brain was quickly dashed by critiques such as Crick (1989), who argued that there was little neurobiological evidence for back-propagation of error in the brain. However, the notion of error propagation is now a part of many influential theories such as predictive coding (Rao & Ballard, 1999; Friston, 2005) and neuroimaging support for these theories is growing (Kok, Rahnev, Jehee, Lau, & de Lange, 2011; Wacongne et al., 2011). Although Error Propagation is not a complete account of the large literature on ERPs, if its central claims are correct, then, by inference to the best explanation, it argues that ERPs are electrical evidence that the brain implements error-based learning algorithms that are functionally equivalent to back-propagation of error (Marblestone, Wayne, & Kording, 2016; Whittington & Bogacz, 2019). In this way, learning by error propagation may be a key part of explaining how the electrochemical brain implements the predictive mind.

Acknowledgments

Hartmut Fitz is a member of the Donders Institute for Brain, Cognition and Behavior Nijmegen and the Max Planck Institute for Psycholinguistics Nijmegen. This work was partially funded by the Netherlands Organisation for Scientific Research (NWO) Gravitation Grant 024.001.006 to the *Language in Interaction* Consortium. Franklin Chang is a member of the International Centre for Language and Communicative Development (LuCiD). The support of the Economic and Social Research Council [ES/L008955/1] is gratefully acknowledged. Correspondence should be sent to Franklin Chang, Kobe City University of Foreign Studies, 9 Chome-1 Gakuen Higashimachi, Nishi Ward, Kobe, Hyogo Prefecture, 651-2187, JAPAN. E-mail: fchang@inst.kobe-cufs.ac.jp.

Appendix A. Simulation details

Simulations were run in the LENS connectionist software package (Rohde et al., 1999) ported to OSX-1.0a (Brouwer, de Kok, & Fitz, 2013). Unless otherwise stated, default parameters of the simulator were used. The code for the simulations is available at: https://sites.google.com/site/sentenceproductionmodel/Home/erpmodel.

The model's sequencing system mapped from the PrevWord layer (88 units) to the Hidden layer (50 units). There was a PrevWordHistory layer (88 units), which received copy connections from itself and the PrevWord layer. This means its activation was a running summary of the PrevWord activation with the most recent words being more activated than earlier words. The Hidden layer was connected to the NextWord layer (88 units) through a Compress layer (30 units). A Context layer (50 units) held a copy of the Hidden layer activation at the previous time step and was fully connected to the Hidden layer. At the start of each utterance, all Context units were reset to 0.5. The NextWord layer used the *soft-max* activation function to create a continuous winner-take-all bias for that layer. The PrevWord layer received one-to-one inputs from all of the NextWord layer units and from the previous target outputs, and a winner-take-all filter was applied. Thus, during learning from the speech of others, the PrevWord was set to the sum of the overheard target word and the model's own internal word predictions.

Messages were stored in the weights between the Role-Concept bindings (Fig. 9, top panel, left), which consisted of the Role layer (6 units) and the Concept layer (62 units). The Hidden layer connected to the Role layer, which connected to the Concept layer. The Concept layer, in turn, connected to the NextWord layer. The weights between the Role and Concept layers were initially cleared, then for a particular message these Role-Concept bindings between appropriate units (e.g., Agent = DOG) were set to a weight of 6 and these weights did not change with learning. To allow the model to recognize the role of previously produced words, the model employed a comprehension message (Fig. 9, top panel, left). This was identical to the production message, except the direction was reversed, mapping from concepts to roles, via weights between the CConcept (62 units) and CRole (6 units) layers. It is assumed that non-linguistic meaning is used to set the CRole-CConcept and Role-Concept bindings simultaneously.

The PrevWord layer connected to the CConcept layer, which connected to the CRole layer, which in turn connected to the Hidden layer (Fig. 9, top panel, left). To ensure the model could avoid producing roles that had already been produced, there was also a CRoleHistory layer (Fig. 9, top panel, center) which averaged a copy of its own activation with the previous activation of the CRole layer. To learn the links between the previous word and its appropriate concept (i.e., the weights between the PrevWord and CConcept layers, Fig. 1, top panel, left), the previous activation of the Concept layer was used as a training signal for the CConcept layer (light grey line, Fig. 9). Finally, the meaning system also included an EventSemantics layer (22 units) connected to the Hidden layer. The Role-Concept links in the production message, the CConcept-CRole links in the comprehension message, and EventSemantics activations were all set before a training or test sentence was processed. Unless specified otherwise, units in all layers used the logistic activation function, with activation values running between 0 and 1. Weights were initially set to values uniformly sampled between -1 and 1. Units were unbiased in order to make layers more dependent on their inputs for their behavior. However, Concept and CConcept units were biased to -3 to ensure that they had a low default activation level.

Steepest descent back-propagation was used. Weights were updated after each message-sentence pair had been trained; the term *epoch* therefore refers to the time taken to train one message-sentence pair. The learning rate was 0.1 throughout training. Training ended after 100,000 sentences had been processed.

Prediction error is a central concept in the proposed account of sentence-level ERP components. Here we explain how it was calculated for different layers in the Dual-path model. Let o_j with $j \in \{1, \ldots, n\}$ be the NextWord layer output units. For these units, the soft-max transfer function was used

$$y_j = \frac{e^{z_j}}{\sum_{i=1}^{n} e^{z_i}} \tag{A.1}$$

where z_j is the net input to unit o_j and y_j its activation output. The exponential magnifies differences in the net inputs and the result is normalized. This continuous winner-take-all function is appropriate for multinomial classification because outputs can be interpreted as a probability distribution over words. A natural match for soft-max is the *divergence* error function

$$E = \sum_{i=1}^{n} t_i \ln\left(\frac{t_i}{y_i}\right) \tag{A.2}$$

where $t_i \in \{0, 1\}$ is the binary target output value of unit o_i. For each NextWord unit o_j, prediction error is measured as the derivative of E with respect to the unit's net input z_j which equals

$$\frac{\partial E}{\partial z_j} = \sum_{i=1}^{n} \left(\frac{\partial E}{\partial y_i}\frac{\partial y_i}{\partial z_j}\right) \tag{A.3}$$

by the chain rule. We now determine the two partial derivatives on the right hand side in (A.3). Clearly,

$$\frac{\partial E}{\partial y_i} = \frac{\partial \sum_{i=1}^{n} t_i(\ln t_i - \ln y_i)}{\partial y_i} = -\frac{t_i}{y_i}. \tag{A.4}$$

To obtain the soft-max derivative, two cases need to be distinguished.

Case 1: $i = j$

$$\frac{\partial y_i}{\partial z_j} = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} - \frac{e^{z_i} e^{z_j}}{(\sum_{j=1}^{n} e^{z_j})^2} = y_i(1 - y_j). \tag{A.5}$$

Case 2: $i \neq j$

$$\frac{\partial y_i}{\partial z_j} = -\frac{e^{z_i} e^{z_j}}{(\sum_{j=1}^{n} e^{z_j})^2} = -y_i y_j. \tag{A.6}$$

Consequently,

$$\frac{\partial y_i}{\partial z_j} = y_i(\delta_{ij} - y_j) \tag{A.7}$$

where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. Substituting (A.4) and (A.7) in (A.3) yields

$$\frac{\partial E}{\partial z_j} = -\sum_{i=1}^{n} t_i(\delta_{ij} - y_j) = \sum_{i=1}^{n} t_i y_j - \sum_{i=1}^{n} t_i \delta_{ij}. \tag{A.8}$$

Because $\sum_{i=1}^{n} t_i = 1$ and $\delta_{ij} = 1$ for $i = j$ only, we obtain

$$\frac{\partial E}{\partial z_j} = y_j - t_j. \tag{A.9}$$

Thus at the NextWord layer, prediction error of unit o_j is simply the difference between the unit's activation and the target value. A positive derivative indicates that a non-target word was predicted whereas a negative derivative indicates that a target word was not fully predicted.

Now, let o_k with $k \in \{1, ..., m\}$ be the m Compress layer units, then

$$\frac{\partial E}{\partial z_k} = \sum_{j=1}^{n} \left(\frac{\partial E}{\partial z_j} \frac{\partial z_j}{\partial y_k} \frac{\partial y_k}{\partial z_k} \right) \tag{A.10}$$

where the sum runs over all NextWord layer units o_j. Since every unit o_j receives input from each Compress unit o_k, we have $z_j = \sum_{k=1}^{m} y_k w_{kj}$ where w_{kj} is the weight from unit k to unit j. Hence

$$\frac{\partial z_j}{\partial y_k} = w_{kj}. \tag{A.11}$$

For the logistic activation function $y_k = \frac{1}{1 + e^{-z_k}}$ that was used for all internal network units we have

$$\frac{\partial y_k}{\partial z_k} = \frac{1}{1 + e^{-z_k}} - \frac{e^{2z_k}}{(1 + e^{z_k})^2} = y_k(1 - y_k). \tag{A.12}$$

Let $\delta_j := \frac{\partial E}{\partial z_j}$ and substitute together with (A.11) and (A.12) in (A.10) to obtain the partial derivative of the error for Compress unit k with respect to its net input

$$\delta_k := \frac{\partial E}{\partial z_k} = y_k(1 - y_k) \sum_{j=1}^{n} \delta_j w_{kj}. \tag{A.13}$$

Error derivatives for deeper layers in the network such as the Hidden layer were determined in the same way. In the back-propagation algorithm these derivatives are used to adjust the weights between layers (Rumelhart et al., 1986). We used them instead to obtain an aggregate measure of prediction error by taking the sum of the absolute value of these deltas for each layer.

Training began by randomizing all weights and the same seed was used for all runs. Model subjects differed in terms of the set of training items they were exposed to. At the start of each utterance, the message was set and did not change throughout production. After the sentence was generated, the sequence of NextWord activations was processed by a decoder program that yielded the produced sentence. Sentences were then processed by a syntactic coder program that added the syntactic and message tags. The model's output was compared with the target sentence and an utterance was considered accurate if all the words and inflectional morphemes were correctly produced.

Appendix B. Grammar that generated model input

The model was trained on message-sentence pairs that were generated by a symbolic grammar. The symbolic grammar had various action categories for verbs: UNACCUSATIVE (4), UNERGATIVE (4), TRANSITIVE (7), BELIEVE (2), DATIVE (4). There were 18 LIVING concepts and 15 NONLIVING concepts (ENTITY included both LIVING and NONLIVING concepts). The category of

NUMber was set so that SINGular was three times more likely than PLURal. DETerminer category was set so that DEFinite was twice as likely as INDEFinite and PROnomial occurred the remaining 14% of the time. There were three TENSE concepts with PRESent and PAST equally frequent and FUTURE occurred the remaining 14% of the time. SIMPle ASPECT was twice as common as PROGressive. Table B.1 is similar to Table 3, except that for each construction its input proportion is shown, as well as the categories in the symbolic grammar. For example to generate the phrase *the toy -s* from 0Y = NONLIVING,DET,NUM, the system would randomly select one of the 15 non-living concepts (e.g., *toy*). Then it would select one of the determiner categories (e.g., DEF) and one of the number categories (e.g., PLUR). Then English rules of syntax would order these elements.

Table B.1
Grammar for generating message-sentence pairs.

Type	Prop.	Example Message-Sentence Pair
Unaccusative Intransitive	.31	0A = UNACCUSATIVE 0Y = NONLIVING,DET,NUM 0E = TENSE,ASPECT,AA,YY *the toy -s were bounce -ing.*
Unergative Intransitive	.04	0A = UNERGATIVE 0Y = LIVING,DET,NUM 0E = TENSE,ASPECT,AA,YY *the grandma -s walk -ed.*
Unergative Locative	.04	0A = UNERGATIVE 0Y = LIVING,THE 1Y = ENTITY,DET,NUM,PREP 0E = TENSE,ASPECT,AA,YY *the sister jump -ed near a husband.*
Active Transitive	0.42	0A = TRANSITIVE 0X = LIVING,DET,NUM 0Y = NONLIVING,DET,NUM 0E = TENSE,ASPECT,AA,XX,YY *the teacher will sniff the wine -s.*
Passive Transitive	.04	0A = TRANSITIVE 0X = LIVING,DET,NUM 0Y = NONLIVING,DET,NUM 0E = TENSE,ASPECT,AA,-3,XX,-10,YY *it will be push -par by the girl.*
Believe Transitive	.04	0A = BELIEVE 0X = LIVING,DET,NUM 0Y = LIVING,DET,NUM 0E = TENSE,ASPECT,AA,XX,YY *she will believe the father.*
Believe Sentence Complement	.04	0A = BELIEVE 0X = LIVING,DET,NUM 0E = TENSE,ASPECT,AA,XX,YY 1A = UNERGATIVE 1Y = LIVING,DET,NUM 1E = TENSE,ASPECT *the man -s will believe that he walk -ed.*
Prepositional Dative	.04	0A = DATIVE 0X = LIVING,DET,NUM 0Y = NONLIVING,DET,NUM 0Z = LIVING,DET,NUM 0E = TENSE,ASPECT,AA,XX,ZZ,YY *a father will send the beer -s to them.*
Double Object Dative	.04	0A = DATIVE 0X = LIVING,DET,NUM 0Y = NONLIVING,DET,NUM 0Z = LIVING,DET,NUM 0E = TENSE,ASPECT,AA,XX,-3,YY,-10,ZZ *the brother will send him a coffee.*

To create dependencies that were tested in the ERP studies, there were some restrictions on the distribution of arguments with transitive verbs. There were nine transitive verbs: *push, take, drink, eat, sip, sniff, taste, believe,* and *know*. The verb *drink*, *sip, sniff,* and *taste* could occur only with the five drinks *water, coffee, wine, beer,* and *tea*. With the verb *drink*, water was the patient 60% of the time and the remaining 40% was split across the other four drinks. The verb *sip* was likewise paired with *tea* 60% of the time. The verb *sniff* and *taste* were paired with *wine* 40% of the time. To allow verbs to be used as nouns (e.g., *take a nap*), the verb *take* occurred with the intransitive verbs *nap, walk, run,* and *jump* 36% of the time (the remaining arguments were inanimate nouns). The verbs *believe* and *know* occurred in transitive and sentence complement frames equally often.

Ten sets of 50,000 randomly generated training message-sentence pairs were created using a different random seed. These sets were used to train ten model subjects. Each model subject was tested on stimuli that simulated the results from the twelve different ERP studies. Each study had 30 items that were generated from the language and these were in the control condition. Then an additional 30 violation items were created by changing the control items in a way that matched the manipulation in the study (see main text for details). Thus, there were 30 matched control and violation items for each study. As the model was tested, training was turned on so that steepest descent error back-propagation was performed in the same way as in training. Except for the adaptation simulations, the model's weights at the end of training were reloaded before each test item, so that the starting point was the same for each test item. As each word was processed, the model's target, output activation, and input derivatives were recorded for each layer. The absolute value of the input derivatives was computed and summed, which provides the layer Sum Abs. Error for each word in each sentence within each model subject. For each ERP study, we extracted the items corresponding to the critical point where the stimuli diverged and aggregated the data by condition and layer.

Appendix C. Examination of Input-derivatives

In Section 2.2, we argued that mechanism behind the N400 and the P600 in the model had different properties. These differences stem from the differences in Eq. (1) for derivative at the NextWord layer and (2) for the derivative at the Hidden layer. It is useful to confirm that our simulations in fact conform to the behavior expected from these equations. To do this, we plotted the unit activation values against the absolute value error derivatives for each unit in the NextWord layer for values from the Cloze study in Section 2.3 (Fig. C.1). Since there are 88 NextWord units, we cannot show all of the items, so we show four items for each of the three conditions,

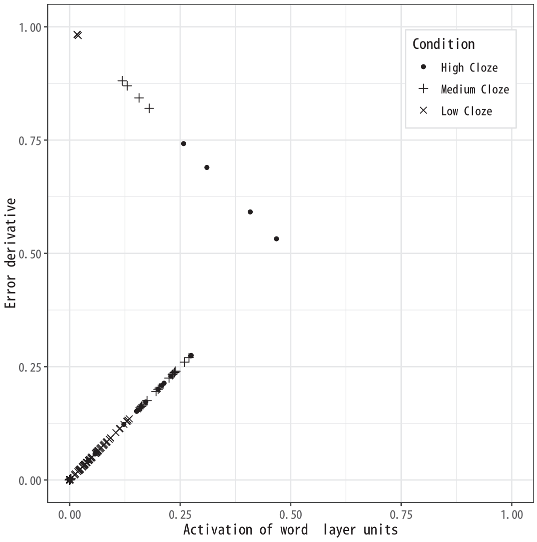

Fig. C.1. Input derivative by activation for NextWord layer in Cloze study.

which allows us to see of some of the variation in the Cloze values. The target pattern has a 1 for actual next word and 0 for all other words. For the actual next word unit, the derivatives are between 0.5 and 1. The Low Cloze data are not expected in this context and have a low activation, so when they are the targets, there is a large absolute value error near 1. As the activation becomes bigger with the High Cloze items, the absolute value error is reduced, because the error is simply the activation of the unit minus the target value of 1 (Eq. 1). In the bottom half of the figure are data points for the many non-target words units. These points have a low activation, because they are often not expected in this context. Since the Cloze manipulation does not influence these points, there is no clear separation in the points as with the target words in the top half of the figure. For words other than the actual next word, the target is 0 and hence the error is just the same as their activation. Since the next NextWord layer activation is constrained by the soft-max activation function to sum to 1, the activation of that layer encodes cloze probability. Since the non-absolute value error is directly related to activation, there is a correlation of -0.999 between error and the cloze probability of that word and hence it is unlikely that another factor will be a better predictor of N400 than production cloze probability in this model. In humans, DeLong et al. (2005) report a correlation between target word cloze and the vertex of the N400 of -0.79. The model's correlation is stronger than the human correlation, because the human N400 includes noise between the source and the scalp that is not included in the model.

To better understand the P600, we create a similar derivative by activation figure for all of the units in the Hidden layer for four items in both condition in the Tense Inflection study in Section 2.7. In this Fig. C.2, there is no correlation between predicted activation and absolute value error derivatives. Instead we see that across a range of difference activation values, control items show smaller error derivatives and violation items show higher error derivatives. This confirms the pattern in the middle panel of Fig. 6,

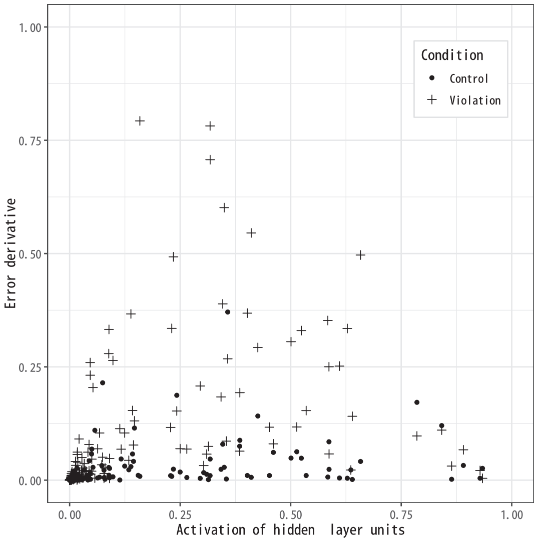

Fig. C.2. Input derivative by activation for Hidden Layer in Tense Inflection study.

where S1 and S2 both have lower magnitude weighted error values in the Grammatical/Control condition and larger magnitude weighted error values in the Ungrammatical/Violation condition. S1 approximates one of the highly activated Hidden units in Fig. C.2, and S2 approximates one of the less activated units. So regardless of the activation of the Hidden units, the weighted error is larger in the ungrammatical than grammatical condition and this creates the P600 in the model.

Inspection of these images shows that the derivatives that support the N400 and P600 have different relationships to predicted activation that is consistent with the different Eqs. (1) and (2) for NextWord and Hidden layers within back-propagation of error.

References

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

Allen, M., Badecker, W., & Osterhout, L. (2003). Morphological analysis in sentence processing: An ERP study. *Language and Cognitive Processes, 18*, 405–430.

Ashby, J., Rayner, K., & Clifton, C. (2005). Eye movements of highly skilled and average readers: Differential effects of frequency and predictability. *The Quarterly Journal of Experimental Psychology Section A, 58*, 1065–1086. https://doi.org/10.1080/02724980443000476.

Atchley, R. A., Rice, M. L., Betz, S. K., Kwasny, K. M., Sereno, J. A., & Jongman, A. (2006). A comparison of semantic and syntactic event related potentials generated by children and adults. *Brain and Language, 99*, 236–246. https://doi.org/10.1016/j.bandl.2005.08.005.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278. https://doi.org/10.1016/j.jml.2012.11.001.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48. https://doi.org/10.18637/jss.v067.i01.

Besson, M., Faita, F., Czternasty, C., & Kutas, M. (1997). What's in a pause: Event-related potential analysis of temporal disruptions in written and spoken sentences. *Biological Psychiatry, 46*, 3–23.

Besson, M., Kutas, M., & Van Petten, C. (1992). An event-related potential (ERP) analysis of semantic congruity and repetition effects in sentences. *Journal of Cognitive Neuroscience, 4*, 132–149.

Bock, K. (1986). Syntactic persistence in language production. *Cognitive Psychology, 18*, 355–387.

Bock, K., Dell, G. S., Chang, F., & Onishi, K. H. (2007). Persistent structural priming from language comprehension to language production. *Cognition, 104*, 437–458.

Bock, K., & Griffin, Z. M. (2000). The persistence of structural priming: Transient activation or implicit learning? *Journal of Experimental Psychology: General, 129*, 177–192.

Bock, K., & Loebell, H. (1990). Framing sentences. *Cognition, 35*, 1–39.

Bornkessel, I., & Schlesewsky, M. (2006). The extended argument dependency model: A neurocognitive approach to sentence comprehension across languages. *Psychological Review, 113*, 787–821. https://doi.org/10.1037/0033-295X.113.4.787.

Borovsky, A., Elman, J. L., & Kutas, M. (2012). Once is enough: N400 indexes semantic integration of novel word meanings from a single exposure in context. *Language Learning and Development, 8*, 278–302.

Borovsky, A., Kutas, M., & Elman, J. (2010). Learning to use words: Event-related potentials index single-shot contextual word learning. *Cognition, 116*, 289–296.

Branigan, H. P., & Messenger, K. (2016). Consistent and cumulative effects of syntactic experience in children's sentence production: Evidence for error-based implicit learning. *Cognition, 157*, 250–256.

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., de Kok, D., & Fitz, H. (2013). LensOSX [2013]. Retrieved from http://hbrouwer.github.io/lensosx/.

Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. *Cognitive Science, 26*, 609–651.

Chang, F. (2009). Learning to order words: A connectionist model of heavy NP shift and accessibility effects in Japanese and English. *Journal of Memory and Language, 61*, 374–397.

Chang, F., Baumann, M., Pappert, S., & Fitz, H. (2015). Do lemmas speak German? A verb position effect in German structural priming. *Cognitive Science, 39*, 1113–1130.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review, 113*, 234–272.

Chang, F., Janciauskas, M., & Fitz, H. (2012). Language adaptation and learning: Getting explicit about implicit learning. *Language and Linguistics Compass, 6*, 259–278.

Cheyette, S. J., & Plaut, D. C. (2017). Modeling the N400 ERP component as transient semantic over-activation within a neural network model of word comprehension. *Cognition, 162*, 153–166.

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*, 1–26. https://doi.org/10.1080/23273798.2018.1427878.

Chow, W.-Y., & Phillips, C. (2013). No semantic illusions in the "Semantic P600" phenomenon: ERP evidence from Mandarin Chinese. *Brain Research, 1506*, 76–93.

Christiansen, M. H., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science, 23*, 157–205.

Christiansen, M. H., Conway, C. M., & Onnis, L. (2012). Similar neural correlates for language and sequential learning: Evidence from event-related brain potentials. *Language and Cognitive Processes, 27*, 231–256. https://doi.org/10.1080/01690965.2011.606666.

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology, 42*, 368–407.

Clahsen, H., Lück, M., & Hahne, A. (2007). How children process over-regularizations: Evidence from event-related brain potentials. *Journal of Child Language, 34*, 601–622. https://doi.org/10.1017/S0305000907008082.

Clifton, C., Traxler, M. J., Mohamed, M. T., Williams, R. S., Morris, R. K., & Rayner, K. (2003). The use of thematic role information in parsing: Syntactic processing autonomy revisited. *Journal of Memory and Language, 49*, 317–334.

Coulson, S., King, J. W., & Kutas, M. (1998). Expect the unexpected: Event-related brain response to morphosyntactic violations. *Language and Cognitive Processes, 13*, 21–58.

Crick, F. (1989). The recent excitement about neural networks. *Nature, 337*, 129–132. https://doi.org/10.1038/337129a0.

Dambacher, M., & Kliegl, R. (2007). Synchronizing timelines: Relations between fixation durations and N400 amplitudes during sentence reading. *Brain Research, 1155*, 147–162. https://doi.org/10.1016/j.brainres.2007.04.027.

Delaney-Busch, N., Morgan, E., Lau, E., & Kuperberg, G. R. (2019). Neural evidence for Bayesian trial-by-trial adaptation on the N400 during semantic priming. *Cognition, 187*, 10–20. https://doi.org/10.1016/j.cognition.2019.01.001.

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences,* (369:20120394), 1–9. https://doi.org/10.1098/rstb.2012.0394.

Dell, G. S., Oppenheim, G. M., & Kittredge, A. K. (2008). Saying the right word at the right time: Syntagmatic and paradigmatic interference in sentence production. *Language and Cognitive Processes, 23*, 583–608. https://doi.org/10.1080/01690960801920735.

Dell, G. S., Reed, K. D., Adams, D. R., & Meyer, A. S. (2000). Speech errors, phonotactic constraints, and implicit learning: A study of the role of experience in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 1355–1367.

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

De Long, K.A., Urbach, T.P., & Kutas, M. (2017). Concerns with Nieuwland et Al. Multi-Lab Study (2017). Technical Report. https://doi.org/10.13140/RG.2.2.19318.

60486.

Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A. M., & Kliegl, R. (2011). Coregistration of eye movements and EEG in natural reading: Analyses and review. *Journal of Experimental Psychology: General, 140*, 552–572. https://doi.org/10.1037/a0023885.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior, 20*, 641–655. https://doi.org/10.1016/S0022-5371(81)90220-6.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179–211.

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

Felser, C., Clahsen, H., & Münte, T. F. (2003). Storage and integration in the processing of filler-gap dependencies: An ERP study of topicalization and wh-movement in German. *Brain and Language, 87*, 345–354.

Fernandez Monsalve, I., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 398–408). Avignon, France.

Ferreira, F., & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language, 25*, 348–368.

Ferreira, F., & Henderson, J. M. (1990). Use of verb information in syntactic parsing: Evidence from eye movements and word-by-word self-paced reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 555–568.

Ferreira, F., & Patson, N. D. (2007). The Good Enough approach to language comprehension. *Language and Linguistics Compass, 1*, 71–83. https://doi.org/10.1111/j.1749-818X.2007.00007.x.

Fiebach, C. J., Schlesewsky, M., & Friederici, A. D. (2002). Separating syntactic memory costs and syntactic integration costs during parsing: The processing of German wh-questions. *Journal of Memory and Language, 47*, 250–272.

Fine, A. B., & Jaeger, T. F. (2013). Evidence for implicit learning in syntactic comprehension. *Cognitive Science, 37*, 578–591. https://doi.org/10.1111/cogs.12022.

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLOS One, 8*, e77661. https://doi.org/10.1371/journal.pone.0077661.

Fischler, I., Bloom, P. A., Childers, D. G., Arroyo, A., & Perry, N. W. (1984). Brain potentials during sentence verification: Late negativity and long-term memory strength. *Neuropsychologia, 22*, 559–568. https://doi.org/10.1016/0028-3932(84)90020-4.

Fitz, H., & Chang, F. (2017). Meaningful questions: The acquisition of auxiliary inversion in a connectionist model of sentence production. *Cognition, 166*, 225–250.

Fitz, H., Chang, F., & Christiansen, M. H. (2011). A connectionist account of the acquisition and processing of relative clauses. In E. Kidd (Ed.), *The Acquisition of Relative Clauses* (pp. 39–60). Amsterdam: John Benjamins.

Foucart, A., & Frenck-Mestre, C. (2012). Can late L2 learners acquire new grammatical features? Evidence from ERPs and eye-tracking. *Journal of Memory and Language, 66*, 226–248. https://doi.org/10.1016/j.jml.2011.07.007.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language, 140*, 1–11. https://doi.org/10.1016/j.bandl.2014.10.006.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology, 14*, 178–210.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences, 6*, 78–84.

Friederici, A. D., & Frisch, S. (2000). Verb argument structure processing: The role of verb-specific and argument-specific information. *Journal of Memory and Language, 43*, 476–507.

Friederici, A. D., Hahne, A., & Mecklinger, A. (1996). Temporal structure of syntactic parsing: Early and late event-related brain potential effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1219–1248.

Friederici, A. D., Hahne, A., & von Cramon, D. Y. (1998). First-pass versus second-pass parsing processes in a Wernicke's and a Broca's aphasic: Electrophysiological evidence for a double dissociation. *Brain and Language, 62*, 311–341.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 360*, 815–836.

Fritsch, F. N., & Carlson, R. E. (1980). Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis, 17*, 238–246.

Garnsey, S. M., Pearlmutter, N. J., Myers, E., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language, 37*, 58–93.

Gehring, W. J., Goss, B., Coles, M. G., Meyer, D. E., & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science, 4*, 385–390.

Gordon, J. K., & Dell, G. S. (2003). Learning to divide the labor: An account of deficits in light and heavy verb production. *Cognitive Science, 27*, 1–40.

Gouvea, A. C., Phillips, C., Kazanina, N., & Poeppel, D. (2010). The linguistic processes underlying the P600. *Language and Cognitive Processes, 25*, 149–188. https://doi.org/10.1080/01690960902965951.

Guillem, F., N'Kaoua, B., Rougier, A., & Claverie, B. (1995). Intracranial topography of event-related potentials (N400/P600) elicited during a continuous recognition memory task. *Psychophysiology, 32*, 382–392. https://doi.org/10.1111/j.1469-8986.1995.tb01221.x.

Hagoort, P., Baggio, G., & Willems, R. (2009). Semantic unification. In M. Gazzaniga (Ed.), *The New Cognitive Neuroscience* (pp. 1–18, 4th ed.). Cambridge, MA: MIT Press.

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

Hagoort, P., Wassenaar, M., & Brown, C. M. (2003). Syntax-related ERP-effects in Dutch. *Cognitive Brain Research, 16*, 38–50.

Hahne, A., Eckstein, K., & Friederici, A. D. (2004). Brain signatures of syntactic and semantic processes during children's language development. *Journal of Cognitive Neuroscience, 16*, 1302–1318.

Harrington Stack, C. M., James, A. N., & Watson, D. G. (2018). A failure to replicate rapid syntactic adaptation in comprehension. *Memory & Cognition, 46*, 864–877. https://doi.org/10.3758/s13421-018-0808-6.

Hoeks, J., Stowe, L., & Doedens, G. (2004). Seeing words in context: The interaction of lexical and sentence level information during reading. *Cognitive Brain Research, 19*, 59–73.

Holroyd, C. B., & Coles, M. G. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review, 109*, 679.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal, 50*, 346–363.

Jaeger, T. F., & Snider, N. E. (2013). Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition, 127*, 57–83. https://doi.org/10.1016/j.cognition.2012.10.013.

Janciauskas, M., & Chang, F. (2018). Input and age-dependent variation in second language learning: A connectionist account. *Cognitive Science, 42*, 519–554. https://doi.org/10.1111/cogs.12519.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*, 122–149.

Kaan, E. (2002). Investigating the effects of distance and number interference in processing subject-verb dependencies: An ERP study. *Journal of Psycholinguistic Research, 31*, 165–193.

Kaan, E. (2007). Event-related potentials and language processing: A brief overview. *Language and Linguistics Compass, 1*, 571–591.

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

Kamide, Y. (2012). Learning individual talkers' structural preferences. *Cognition, 124*, 66–71.

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

Kim, J. (1998). Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation. Cambridge, Mass: MIT press.

Kok, P., Rahnev, D., Jehee, J. F., Lau, H. C., & de Lange, F. P. (2011). Attention reverses the effect of prediction in silencing sensory signals. *Cerebral Cortex, 22*, 2197–2206.

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*, 1–36. https://doi.org/10.1016/S0093-934X(02)00548-5.

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research, 1146*, 23–49. https://doi.org/10.1016/j.brainres.2006.12.063.

Kuperberg, G. R., Caplan, D. N., Sitnikova, T., & Holcomb, P. J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cognitive Brain Research, 17*, 117–129.

Kuperberg, G. R., Kreher, D. N., Caplan, D. A., Sitnikova, T., & Holcomb, P. J. (2007). The role of animacy and thematic relationships in processing active English sentences: Evidence from even-related potentials. *Brain Research, 100*, 223–237.

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

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology, 62*, 621–647.

Kutas, M., & Hillyard, S. A. (1980). Reading between the lines: Event-related brain potentials during natural sentence processing. *Brain and Language, 11*, 354–373.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307*, 161–163.

Laszlo, S., & Armstrong, B. C. (2014). PSPs and ERPs: Applying the dynamics of post-synaptic potentials to individual units in simulation of temporally extended event-related potential reading data. *Brain and Language, 132*, 22–27. https://doi.org/10.1016/j.bandl.2014.03.002.

Laszlo, S., & Federmeier, K. D. (2011). The N400 as a snapshot of interactive processing: Evidence from regression analyses of orthographic neighbor and lexical associate effects. *Psychophysiology, 48*, 176–186. https://doi.org/10.1111/j.1469-8986.2010.01058.x.

Laszlo, S., & Plaut, D. C. (2012). A neurally plausible Parallel Distributed Processing model of event-related potential word reading data. *Brain and Language, 120*, 271–281. https://doi.org/10.1016/j.bandl.2011.09.001.

Lenth, R. (2017). Emmeans: Estimated marginal means, aka least-squares means. https://CRAN.R-project.org/package=emmeans.

Liu, L., Burchill, Z., Tanenhaus, M. K., & Jaeger, T.F. (2017). Failure to replicate talker-specific syntactic adaptation. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 2616–2621). London.

Lück, M., Hahne, A., & Clahsen, H. (2006). Brain potentials to morphologically complex words during listening. *Brain Research, 1077*, 144–152.

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Language Sciences, 4*, 226. https://doi.org/10.3389/fpsyg.2013.00226.

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

Marblestone, A. H., Wayne, G., & Kording, K. P. (2016). Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience, 10*, 1–41.

Martín-Loeches, M., Casado, P., Gonzalo, R., de Heras, L., & Fernández-Frías, C. (2006). Brain potentials to mathematical syntax problems. *Psychophysiology, 43*, 579–591. https://doi.org/10.1111/j.1469-8986.2006.00463.x.

McLaughlin, J., Osterhout, L., & Kim, A. (2004). Neural correlates of second-language word learning: Minimal instruction produces rapid change. *Nature Neuroscience, 7*, 703–704. https://doi.org/10.1038/nn1264.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, 37*, 547–559.

Mestres-Misse, A., Rodriguez-Fornells, A., & Munte, T. F. (2007). Watching the brain during meaning acquisition. *Cerebral Cortex, 17*, 1858–1866. https://doi.org/10.1093/cercor/bhl094.

Moreno, E. M., Federmeier, K. D., & Kutas, M. (2002). Switching languages, switching palabras (words): An electrophysiological study of code switching. *Brain and Language, 80*, 188–207.

Nakano, H., Saron, C., & Swaab, T. Y. (2010). Speech and span: Working memory capacity impacts the use of animacy but not of world knowledge during spoken sentence comprehension. *Journal of Cognitive Neuroscience, 22*, 2886–2898.

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., …, Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*. https://doi.org/10.7554/eLife.33468.

Noppeney, U., & Price, C. J. (2004). An fMRI study of syntactic adaptation. *Journal of Cognitive Neuroscience, 16*, 702–713.

Olichney, J. M., Van Petten, C., Paller, K. A., Salmon, D. P., Iragui, V. J., & Kutas, M. (2000). Word repetition in amnesia: Electrophysiological measures of impaired and spared memory. *Brain, 123*, 1948–1963.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language, 31*, 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*, 786–803.

Osterhout, L., & Mobley, L. A. (1995). Event-related potentials elicited by failure to agree. *Journal of Memory and Language, 34*, 739–773.

Otten, M., & Van Berkum, J. J. (2008). Discourse-based word anticipation during language processing: Prediction or priming? *Discourse Processes, 45*, 464–496.

Özyürek, A., Willems, R. M., Kita, S., & Hagoort, P. (2007). On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *Journal of Cognitive Neuroscience, 19*, 605–616.

Patel, A. D., Gibson, E., Ratner, J., Besson, M., & Holcomb, P. J. (1998). Processing syntactic relations in language and music: An event-related potential study. *Journal of Cognitive Neuroscience, 10*, 717–733. https://doi.org/10.1162/089892998563121.

Perfetti, C. A., Wlotko, E. W., & Hart, L. A. (2005). Word learning and individual differences in word learning reflected in event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 1281–1292. https://doi.org/10.1037/0278-7393.31.6.1281.

Phillips, C., Kazanina, N., & Abada, S. H. (2005). ERP effects of the processing of syntactic long-distance dependencies. *Cognitive Brain Research, 22*, 407–428.

Qian, Z., Garnsey, S., & Christianson, K. (2017). A comparison of online and offline measures of good-enough processing in garden-path sentences. *Language, Cognition and Neuroscience,* 1–28.

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*, 693–705. https://doi.org/10.1038/s41562-018-0406-4.

Rabovsky, M., & McRae, K. (2014). Simulating the N400 ERP component as semantic network error: Insights from a feature-based connectionist attractor model of word meaning. *Cognition, 132*, 68–89. https://doi.org/10.1016/j.cognition.2014.03.010.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*, 79. https://doi.org/10.1038/4580.

Rayner, K., & Clifton, C. (2009). Language processing in reading and speech perception is fast and incremental: Implications for event-related potential research. *Biological Psychology, 80*, 4–9. https://doi.org/10.1016/j.biopsycho.2008.05.002.

Rayner, K., & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review, 3*, 504–509.

Reali, F., & Christiansen, M. H. (2005). Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science, 29*, 1007–1028.

Rohde, D. L. (1999). LENS: The Light, Efficient Network Simulator. Technical Report CMU-CS-99-164 Carnegie Mellon University Pittsburgh, PA.

Rommers, J., & Federmeier, K. D. (2018). Predictability's aftermath: Downstream consequences of word predictability as revealed by repetition effects. *Cortex, 101*, 16–30. https://doi.org/10.1016/j.cortex.2017.12.018.

Rugg, M. D., & Curran, T. (2007). Event-related potentials and recognition memory. *Trends in Cognitive Sciences, 11*, 251–257.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*, 533–536.

Schneider, J. M., Abel, A. D., Ogiela, D. A., Middleton, A. E., & Maguire, M. J. (2016). Developmental differences in beta and theta power during sentence processing. *Developmental Cognitive Neuroscience, 19*, 19–30. https://doi.org/10.1016/j.dcn.2016.01.001.

Segaert, K., Menenti, L., Weber, K., Petersson, K. M., & Hagoort, P. (2012). Shared syntax in language production and language comprehension—An fMRI study. *Cerebral Cortex, 22*, 1662–1670. https://doi.org/10.1093/cercor/bhr249.

Sereno, S., & Rayner, K. (2003). Measuring word recognition in reading: Eye movements and event-related potentials. *Trends in Cognitive Sciences, 7*, 489–493. https://doi.org/10.1016/j.tics.2003.09.010.

Sereno, S. C., Rayner, K., & Posner, M. I. (1998). Establishing a time-line of word recognition: Evidence from eye movements and event-related potentials. *Neuroreport,* *9*, 2195–2200.

Sitnikova, T., Holcomb, P. J., Kiyonaga, K. A., & Kuperberg, G. R. (2008). Two neurocognitive mechanisms of semantic integration during the comprehension of visual real-world events. *Journal of Cognitive Neuroscience, 20*, 2037–2057.

St. John, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence, 46*, 217–257.

Swaab, T. Y., Ledoux, K., Camblin, C. C., & Boudewyn, M. A. (2013). Language-related ERP components. In E.S. Kappenman, & S.J. Luck (Eds.), *The Oxford Handbook of Event-Related Potential Components* (pp. 397–440). New York: Oxford University Press.

Thornhill, D. E., & Van Petten, C. (2012). Lexical versus conceptual anticipation during sentence processing: Frontal positivity and N400 ERP components. *International Journal of Psychophysiology, 83*, 382–392. https://doi.org/10.1016/j.ijpsycho.2011.12.007.

Tooley, K. M., & Bock, K. (2014). On the parity of structural persistence in language production and comprehension. *Cognition, 132*, 101–136. https://doi.org/10.1016/j.cognition.2014.04.002.

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language, 33*, 285–318.

Trueswell, J. C., Tanenhaus, M. K., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 528–553. https://doi.org/10.1037//0278-7393.19.3.528.

Twomey, K. E., Chang, F., & Ambridge, B. (2014). Do as I say, not as I do: A lexical distributional account of English locative verb class acquisition. *Cognitive Psychology, 73*, 41–71. https://doi.org/10.1016/j.cogpsych.2014.05.001.

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*, 443–467.

Van Berkum, J. J., Hagoort, P., & Brown, C. M. (1999). Semantic integration in sentences and discourse: Evidence from the N400. *Journal of Cognitive Neuroscience, 11*, 657–671.

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

Van Petten, C. (1993). A comparison of lexical and sentence-level context effects in event-related potentials. *Language and Cognitive Processes, 8*, 485–532.

Van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 394–417.

Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Memory & Cognition, 18*, 380–393.

Van Petten, C., & Kutas, M. (1991). Influences of semantic and syntactic context on open-and closed-class words. *Memory & Cognition, 19*, 95–112.

Van Petten, C., & Luka, B. L. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology, 83*, 176–190. https://doi.org/10.1016/j.ijpsycho.2011.09.01.

Van Petten, C., Weckerly, J., McIsaac, H. K., & Kutas, M. (1997). Working memory capacity dissociates lexical and sentential context effects. *Psychological Science, 8*, 238–242.

Wacongne, C., Labyt, E., van Wassenhove, V., Bekinschtein, T., Naccache, L., & Dehaene, S. (2011). Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proceedings of the National Academy of Sciences, 108*, 20754–20759. https://doi.org/10.1073/pnas.1117807108.

Wassenaar, M., & Hagoort, P. (2005). Word-category violations in patients with Broca's aphasia: An ERP study. *Brain and Language, 92*, 117–137.

Weber, K., & Lavric, A. (2008). Syntactic anomaly elicits a lexico-semantic (N400) ERP effect in the second language but not the first. *Psychophysiology, 45*, 920–925. https://doi.org/10.1111/j.1469-8986.2008.00691.x.

Whittington, J. C., & Bogacz, R. (2019). Theories of error back-propagation in the brain. *Trends in Cognitive Sciences, 23*, 235–250. https://doi.org/10.1016/j.tics.2018.12.005.

Yan, S., Kuperberg, G. R., & Jaeger, T. F. (2017). Prediction (or not) during language processing. A Commentary On Nieuwland et al. (2017) And Delong et al. (2005). https://doi.org/10.1101/143750.