Cognition 162 (2017) 153–166

Contents lists available at ScienceDirect

Cognition

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

Original Articles

Modeling the N400 ERP component as transient semantic over-activation within a neural network model of word comprehension

Samuel J. Cheyette [a], David C. Plaut [b,*]

[a] Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627, USA
[b] Department of Psychology and the Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA 15213, USA

ARTICLE INFO

Article history:
Received 18 September 2015
Revised 21 October 2016
Accepted 27 October 2016
Available online 18 November 2016

Keywords:
N400
Event-related potentials
Neural networks
Word comprehension

ABSTRACT

The study of the N400 event-related brain potential has provided fundamental insights into the nature of real-time comprehension processes, and its amplitude is modulated by a wide variety of stimulus and context factors. It is generally thought to reflect the difficulty of semantic access, but formulating a precise characterization of this process has proved difficult. Laszlo and colleagues (Laszlo & Plaut, 2012; Laszlo & Armstrong, 2014) used physiologically constrained neural networks to model the N400 as transient over-activation within semantic representations, arising as a consequence of the distribution of excitation and inhibition within and between cortical areas. The current work extends this approach to successfully model effects on both N400 amplitudes and behavior of word frequency, semantic richness, repetition, semantic and associative priming, and orthographic neighborhood size. The account is argued to be preferable to one based on "implicit semantic prediction error" (Rabovsky & McRae, 2014) for a number of reasons, the most fundamental of which is that the current model actually produces N400-like waveforms in its real-time activation dynamics.

1. Introduction

The N400 is a negative deflection in event-related brain potentials (ERPs) that occurs approximately 400 ms post-stimulus onset in response to a wide range of meaningful or potentially meaningful stimuli, including written and spoken words and pseudowords, and drawings, photos and videos of objects and actions (for reviews, see Federmeier & Laszlo, 2009; Kutas & Federmeier, 2009, 2011). It was originally identified as a response to semantically anomalous sentence endings (e.g., "I take coffee with cream and dog"; Kutas & Hillyard, 1980) but, over the years, has been shown to be sensitive to a wide variety of stimulus and context manipulations, including cloze probability (the number of possible sentence endings), sentence and discourse congruity, repetition, semantic priming, lexical association, concreteness and semantic richness, word frequency, orthographic neighborhood size, and many more. On the other hand, N400 amplitude is relatively insensitive to manipulations that broadly preserve meaning, including physical changes (e.g., in font or case) and syntactic violations (e.g., in number agreement). Understanding the N400 is important because it offers a real-time measure linking underlying neural mechanisms to behavior that has provided fundamental insights into core issues in the study of cognitive and neural processing, including the immediacy and incrementality of comprehension, the integration of bottom-up and top-down sources of information, the organization and dynamics of semantic memory, and the bases for variability and atypicality in performance across individuals and in special populations (Kutas & Federmeier, 2011).

The wide range of factors that modulate the N400 is, unfortunately, matched by an equally wide range of theoretical accounts of the phenomena. One proposal is that the N400 reflects post-lexical semantic integration or unification, linking semantic information from a current word with meaningful information from previous words and context (Brown & Hagoort, 1993; Hagoort, Baggio, & Willems, 2009). This broad theory accounts for the N400's largely meaning-specific modulation, but fails to account for many of its subtleties. For instance, it is unclear why an N400 is generated by words in isolation, or even by pseudowords (Deacon, Dynowska, Ritter, & Grose-Fifver, 2009; Laszlo & Federmeier, 2007, 2011), and why its amplitude is modulated by form-based properties such as orthographic neighborhood size (Laszlo & Federmeier, 2009). Other researchers (Deacon, Dynowska, Ritter, & Grose-Fifer, 2004) have suggested that the N400 reflects orthographic/phonological analysis that is attenuated by top-down semantic feedback. In complementary fashion, this account explains sensitivity to lexical and form-based factors

* Corresponding author.
E-mail addresses: sjcheyette@gmail.com (S.J. Cheyette), plaut@cmu.edu (D.C. Plaut).

http://dx.doi.org/10.1016/j.cognition.2016.10.016
0010-0277/

but provides a less satisfactory account of sentence- and discourse-level effects (see van Berkum, 2009).

Perhaps the most common perspective falls between these two extremes: that the N400 reflects something like the difficulty of semantic access (Kutas & Federmeier, 2000, 2011). This proposal is supported in part by attempts to localize the neural generators of the N400 component (e.g., Halgren et al., 2002; Lau, Phillips, & Poeppel, 2008; Van Petten & Luka, 2006), which generally implicate brain regions involved in semantic processing, including the superior/middle temporal gyrus, the temporal-parietal junction, and the medial temporal lobe. It has, however, proved difficult to formulate a precise characterization of "semantic access" that is capable of accounting for the full range of empirical effects. Indeed, in an attempted synthesis from this perspective, Kutas and Federmeier (2011) offered only a very general characterization:

> Rather than reflecting the activation of "a word's meaning," then, the N400 region of the ERP is more accurately described as reflecting the activity in a multimodal long-term memory system that is induced by a given input stimulus during a delimited time window as meaning is dynamically constructed. (p. 640)

Laszlo and Plaut (2012) put forth a specific proposal for the basis of the N400 and supported their account with an explicit computational simulation using a neurophysiologically constrained neural network. On their view, the N400 does, in fact, reflect the activation of a word's meaning, but this process is sensitive to a variety of bottom-up and top-down influences and also exhibits specific temporal dynamics due to the organization of excitation and inhibition within cortex. In particular, it is well established that the projections of pyramidal cells between cortical areas are restricted to be excitatory, whereas inhibitory interneurons operate locally to modulate overall activity levels within each area (Kandel & Schwartz, 1985). As a result, bottom-up input causes an initial over-activation of neurons within an area, which is subsequently resolved into a coherent representation with lower activation through competitive (and cooperative) interactions (see Zheng et al., 2012). Laszlo and Plaut proposed that the N400 deflection reflects this transient over-activation of neurons in cortical areas representing word meaning, and that its magnitude depends both on the nature of co-activated information due to the similarity structure of word forms and meanings, as well as on pre-activated information from prior context.

To support this account, Laszlo and Plaut (2012) developed a neural network simulation of word comprehension that incorporated the relevant constraints on excitation and inhibition between and within layers. In the model, visual input mapped to orthographic, hidden, and semantic representations in turn. At each layer, excitatory units received positive-only bottom-up input from the layer below, and projected positive-only connections to the next layer as well as to an inhibitory unit which projected back with negative-only connections. The model was trained to reconstruct the visual input and to generate the semantic representations for 62 CVC words as well as 15 acronyms (containing a central consonant). Acronyms were included in order to model single-item ERP data gathered by Laszlo and Federmeier (2011) in which they independently varied meaningfulness and orthographic regularity by comparing words (e.g., HAT), pseudowords (e.g., KOF), acronyms (e.g., DVD), and illegal strings (e.g., NHK). Somewhat surprisingly, Laszlo and Federmeier found that N400 magnitude depended on orthographic regularity but not on meaningfulness (see Fig. 1a). Moreover, across all stimulus types, there was a strong correlation between N400 amplitude and orthographic neighborhood size, regardless of lexical status. These results are particularly important because they would seem to be at odds with accounts of the N400 as reflecting semantic access per se.

As it turns out, however, the Laszlo and Plaut (2012) comprehension model shows the same pattern of performance when tested on analogous stimuli (see Fig. 1b). Laszlo and Plaut measured mean semantic activation over time as a proxy for the population-based post-synaptic potentials thought to underlie EEG signals in general, and the N400 component in particular (see Fabiani, Gratton, & Federmeier, 2007). Although the model ultimately settles to stronger semantic representations for meaningful stimuli (words and acronyms), it produces greater transient semantic activation—and, hence, greater N400 amplitudes—for orthographically regular stimuli (words and pseudowords). The reason is that orthographic forms provide bottom-up excitation not only for their specific semantic features but also for the semantic features of orthographically similar forms. Thus, words and pseudowords, with many orthographic neighbors, generate much greater transient semantic over-activation than do acronyms and illegal strings, with few if any neighbors. Laszlo and Plaut showed that the separation of excitation and inhibition is essential to producing these dynamics; an otherwise equivalent but unconstrained network failed to exhibit the empirically observed pattern. In this way, the model provides a specific, neurally explicit instantiation of comprehension processes in which the N400 can be understood as reflecting "semantic access", and yet can nonetheless explain why it occurs as strongly for pseudowords as for words, and why its amplitude depends on form-based properties rather than on meaningfulness. In follow-up simulations, Laszlo and Armstrong (2014) added a fatigue function to the

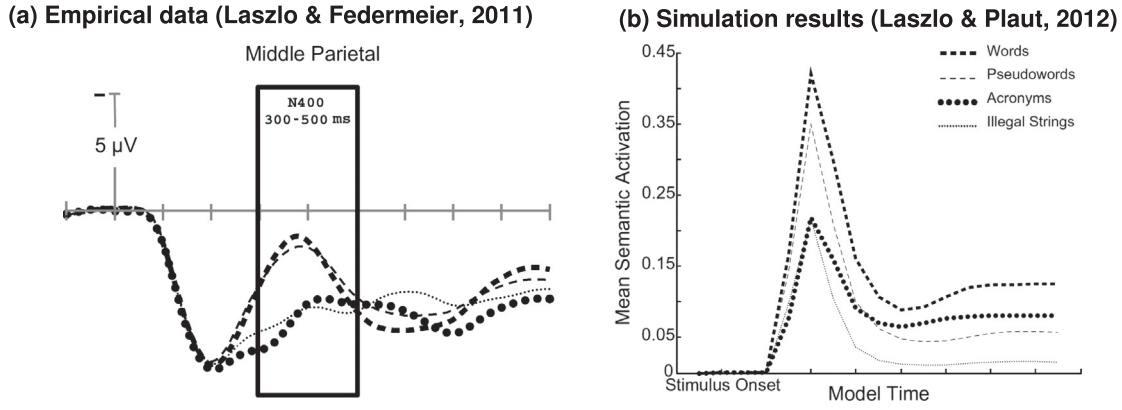

Fig. 1. (a) Empirical data from Laszlo and Federmeier (2011) showing N400 magnitudes to words, pseudowords, acronyms, and illegal strings; (b) Mean semantic activation over time exhibited by the Laszlo and Plaut (2012) simulation for the same stimulus classes. (Reprinted with permission from Laszlo and Plaut, 2012.)

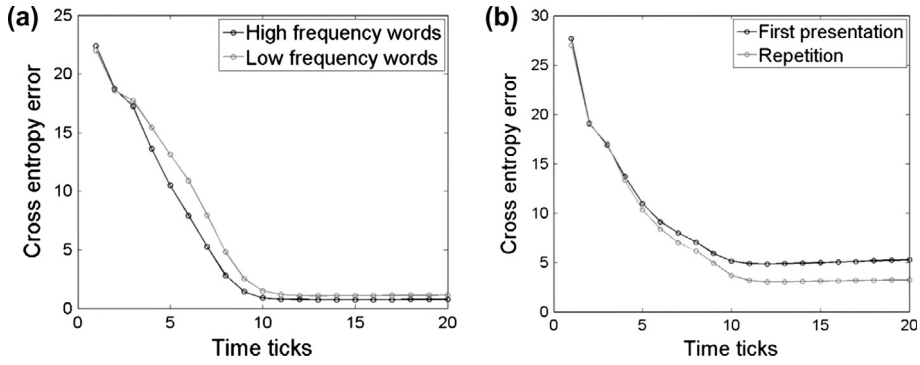

Fig. 2. Rabovsky and McRae's (2014) results using cross-entropy error to simulate the effects on N400 amplitude of (a) word frequency and (b) repetition. (Reprinted with permission from Rabovsky and McRae, 2014.)

excitatory units to account for the reduction in N400 amplitude due to stimulus repetition (e.g., Nagy & Rugg, 1989), which can be viewed as a first step toward accounting for effects of prior context.

Recently, Rabovsky and McRae (2014) proposed an alternative account of the N400 based on a different type of neural network simulation, and applied it to a much broader range of phenomena than addressed by Laszlo and Plaut (2012) and Laszlo and Armstrong (2014). On their account, semantic activation forms the basis for task performance (e.g., lexical decision), whereas the N400 reflects "implicit prediction error" over semantics—the discrepancy between the semantic information derived from a stimulus and the information predicted or anticipated from prior context (where this "prediction" is not conscious or explicit). Rabovsky and McRae supported their account using an attractor network model of word comprehension (Cree, McNorgan, & McRae, 2006). The model consists of a fully connected network without hidden units that was trained to map the orthographic forms of 541 words (over 30 units) to their semantic feature representations (over 2526 units, derived from McRae, Cree, Seidenberg, & McNorgan, 2005). Implicit semantic prediction error was operationalized in terms of task performance error—the discrepancy at any point in time between the semantics generated by the network and the correct semantic pattern for the target stimulus.[1]

Rabovsky and McRae did not attempt to reproduce the actual morphology of the N400 deflection. Rather, they considered only the direction of changes in N400 amplitude, and of behavioral performance, in the context of seven empirical effects: semantic priming, semantic richness (i.e., number of semantic features), word frequency, repetition, and the interactions of repetition with richness and with frequency, and orthographic neighborhood size.[2] They showed that, for each of these effects, the error measure is influenced in the same direction as N400 amplitude, whereas overall semantic activation is influenced in the same direction as behavior. For example, Fig. 2 shows how semantic error in their model varies over time as a function of word frequency and repetition.

Although the Rabovsky and McRae (2014) model is impressive in its breadth of coverage of phenomena relating to both N400 amplitude and behavior, a number of aspects of its design and performance are less than satisfactory. First and foremost, of course, is that the model doesn't actually produce N400 morphology. This is a fundamental limitation because the N400 is not simply a single-valued dependent measure like accuracy or reaction time—it reflects the moment-to-moment changes in activity (post-synaptic potentials) of neural populations that directly contribute to performance, and thus provides a wealth of information linking brain processes to behavior as they occur in real time. A model that fails to address actual N400 dynamics can provide little insight into these deeper issues.

Second, many of the effects in the Rabovsky and McRae model are very small and/or hold only over somewhat different ranges of processing cycles in the model; while this is less a concern about the behavioral measure, it is a serious issue for an account of the N400, as one of its hallmark characteristics is that its latency is relatively stable (Kutas & Federmeier, 2011). Third, it is questionable how semantic prediction error could actually be computed in cortex, and how it would manifest as neural activity—so as to be measurable by EEG—distinct from semantic activation. The Rabovsky and McRae (2014) simulation computed error using explicit targets for semantic features, but even if such targets had an actual neural instantiation—which seems unlikely (Crick, 1989)—they are not directly available in a standard lexical decision paradigm, nor is there any context from which to derive them. Finally, for words in unrelated contexts, such as an unordered list, it is difficult to understand how the system could make sensible predictions of their semantics, and it is unclear how the notion of semantic prediction error applies in the case of pseudowords, which have no semantics and yet produce N400 amplitudes as large as those for words (Laszlo & Federmeier, 2011).

Given these concerns, it seems preferable to us to account for the relevant empirical phenomena within an approach in which the N400 corresponds directly to (semantic) neural activity. In fact, there are reasons to believe that semantic activation in the Laszlo and Plaut (2012) model would behave in ways similar to what Rabovsky and McRae (2014) claim for implicit semantic prediction error (while avoiding its pitfalls). For example, insofar as prior context—including a preceding prime word—activates information that supports the features of the target, those features may inhibit competing features (of orthographic neighbors), thereby reducing N400 amplitude.

Accordingly, in the current work, we set out to account for the same breadth of phenomena as Rabovsky and McRae (2014), concerning effects on both the N400 and behavior, but using the general approach of Laszlo and Plaut (2012) and Laszlo and Armstrong (2014). In the first simulation, the N400 is again modeled by overall semantic activation within a physiologically constrained neural network. In Simulation 2, we augment this network with a trained

[1] Rabovsky and McRae (2014, p. 70) suggest that the network's activation corresponds to the prediction, and the correct semantic targets (of unspecified origin) correspond to the actual outcome. However, on this view, in order for the real-time value of prediction error to correspond to the real-time value of the N400, the "outcome" (correct semantic targets) must be available at the very start of the generation of the "predictions" (network activations), which seems awkward to us. For this reason, we prefer an interpretation in which prior context provides the prediction that is then compared against the actual semantics generated by the word itself, and this is how our discussion is framed throughout the paper.

[2] The specific patterning of these effects will be considered in detail later, in the context of presenting the corresponding modeling work.

response system similar to Usher and McClelland's (2001) leaky competing integrator model of decision making in order to model behavioral effects (in lexical decision).

2. Simulation 1: N400 effects

Simulations were run using a modified version of the Lens neural network simulator developed by Doug Rohde (http://tedlab.mit.edu/dr/Lens/). The code for the modified simulator and all necessary training and testing files are available for download at http://www.cnbc.cmu.edu/plaut/CheyettePlaut-N400.

2.1. Methods

The simulation had roughly the same design as in Laszlo and Plaut (2012, hereafter LP12), with the following main modifications: (1) a larger vocabulary (to allow for variations in word frequency and semantic richness); (2) a concomitant increased in numbers of excitatory and inhibitory units; (3) the introduction of an activation-based decay function similar to that employed by Laszlo and Armstrong (2014); and (4) the introduction of a response system that makes lexical decisions based on semantic input. For completeness, though, we include all simulation details below.[3] Although the new model differs from these previous ones in a number of detailed respects (e.g., exact ratios of excitatory-to-inhibitory units; specific decay function), it retains the same core theoretical commitments: (a) distributed representations of orthography and semantics, and no localist word units; (b) a separation of excitation and inhibition with connectivity constraints that gives rise to early excitation followed by late inhibition; and (c) a form of neural fatigue driven by sustained activation (see also Gotts & Plaut, 2002).

2.1.1. Stimuli

The network was trained to map visual (orthographic) input to semantic output for 176 words with consonant-vowel-consonant (CVC) structure. Visual input was coded over 24 units (8 per letter) with each of 15 possible letters (10 consonants, 5 vowels) activating 2 of 8 units in each slot. Semantic representations were encoded over 70 semantic units, with words varying in semantic richness: half had 6 features and half had 3 features. This made the semantic representations highly sparse, but still allowed for some degree of overlap (to reflect semantic relatedness). Within each richness level, words also varied in frequency, with half occurring 5 times more often during training than the other half. Visual inputs were assigned to semantic outputs randomly to ensure that there was no systematic relationship between the forms of words and their meanings.

2.1.2. Network architecture

The architecture of the network is depicted in Fig. 3. The bottom layers form an autoencoder which is trained to reconstruct each 24-element visual input pattern via an intermediate group of 24 hidden units (labeled "Orthographic Autoencoder" in the figure). These hidden units then map via another group of 90 hidden units to 70 semantic units. These between-layer connections are constrained to excitatory. Each of these groups of units has a corresponding group of inhibitory units. As in the LP12 model, the hidden and output groups within the orthographic autoencoder each has a single inhibitory unit that receives excitatory connections from its corresponding excitatory layer and sends inhibitory

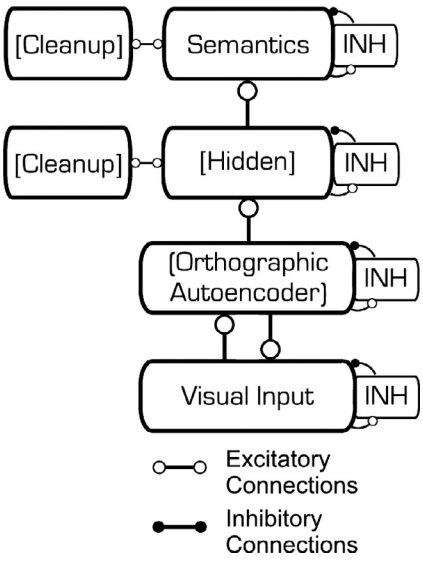

Fig. 3. The architecture of the network used in Simulation 1. Names of hidden layers are in square brackets. The bottom two layers form a feedforward autoencoder with separate input and output layers, but these are depicted as a single "Visual Input" layer with bidirectional connectivity for simplicity.

connections back to it. In order to cope with the larger vocabulary, however, the semantic layer and the hidden layer just below it have 3 inhibitory units each (but with the same connectivity constraints). In addition, they each are bidirectionally connected with positive-only weights to their own set of "clean-up" units (35 for both the hidden and semantic layers) that help the network learn higher-order structure among semantic patterns (Hinton & Shallice, 1991). Including bias connections for all non-input layers, the network has a total of 22,154 connections.

Each excitatory unit computes its activation as the standard logistic (sigmoid) function of its time-averaged net input from other units, which is then subject to multiplicative decay as a function of its time-averaged activation:

$$n_j^t = \tau \sum_i a_i^{t-1} w_{ij} + (1-\tau)\, n_j^{t-1} \quad (1)$$

$$o_j^t = \lambda f\left(n_j^t\right) + (1-\lambda)\, o_j^{t-1} \quad (2)$$

$$a_j^t = f\left(n_j^t\right)\left(1 - o_j^t(1-\beta)\right) \quad (3)$$

where w_{ij} is the weight on the connection from unit i to unit j, n_j^t is the net input of unit j at time t, a_j^t is the instantaneous activation of unit j at time t, o_j^t is the time-averaged activation of unit j at time t, $\tau = 0.5$ is the time constant for averaging net inputs, $\lambda = 0.06$ is the time constant for averaging activations, $\beta = 0.8$ is the upper bound on decay, and $f(x) = 1/(1+e^{-x})$ is the sigmoid function. Note in Eq. 3 that there is no decay if a unit's time-averaged activation o_j^t is 0.0 but full decay of β if the time-averaged activation is 1.0. This type of activation-based decay is simpler than the alpha function used by Laszlo and Armstrong (2014) but has very similar properties.[4] We chose the values of τ, β and λ somewhat arbitrarily, but with the intent that the decay from the peak activation of one word would influence that of the next. Moreover, these parameters give rise to dynamics in which the drop in a unit's activation is relatively rapid after one or two repetitions of a stimulus and then reach asymptote quickly, which agrees with empirical studies of neural repetition suppression (see, e.g., Miller, Gochin, & Gross, 1991).

[3] The findings to be reported are generally stable over small variations in network parameters and over initial random weight values. We report on a single simulation to more clearly convey the network dynamics and behavior of an individual network, which we take to approximate something like a modal participant.

[4] We chose not to employ the alpha function directly because it determines an envelope within which unit activation is constrained, and so—at least in principle—limits a unit's activation both early and late in the course of processing.

Inhibitory units employed the same multi-linear "elbow" function used in LP12 that approximates the combined influence of two inhibitory populations: an immediately active linear population, and a thresholded linear population that becomes active only under stronger net input. The functional consequence of this unit function is that excitatory activation tends to stabilize at a level that is in balance with the amount of inhibition produced at the inflection point (elbow) of the inhibitory function and, in this way, serves as a graded form of k-winner-take-all (see LP12 for further discussion).

2.1.3. Training procedures

As in LP12, the network was trained to take the visual representations of each word as input and to reconstruct this representation via the autoencoder units in Fig. 3. However, unlike LP12, the autoencoder and the rest of the network are trained simultaneously. During this training, words were presented in pairs, with all possible 176 × 176 = 30,976 pairs occurring during training. However, certain pairs had elevated frequencies of occurrence relative to others. In particular, we introduced lexical associations between words in order to address certain aspects of the empirical findings on semantic priming, as will be discussed in detail in the Results section below. Specifically, each word had another word designated at its associate, such that the word was followed by its associate on 30% of its presentations, and by some other word on the remaining 70% of presentations (see also Plaut, 1995; Plaut & Booth, 2000). In addition, the frequency of each word pair was adjusted to enforce the word-specific frequency manipulation that high-frequency words occurred 5 times more often than low-frequency words—this required that words and their associates were matched in frequency.

The timing of presentation of a given word pair was the same as in LP12 in that the input for each word was presented over the Visual Input units for 16 *ticks* (unit updates), with a single tick with zeros as input in between. Semantic targets were applied for the last 12 ticks during the presentation of each word. Unit activations (and the integrated activations that govern decay) were reinitialized between word pairs. The network was trained on 750,000 presentations of word pairs, sampling randomly but according to their specified frequencies of occurrence, using back-propagation for continuous-time networks (Pearlmutter, 1989), cross-entropy error, a batch size of 1, no momentum, and a gradually lowering learning rate: 0.015 for 250,000 presentations, 0.01 for 250,000, and 0.005 for 250,000.[5] The clean-up layers used a reduced learning rate (0.001) for the first 100,000 presentations, as recurrent connections are often not beneficial until some training has occurred (Marr, 1971).

2.1.4. Testing procedures

Following training, we tested the network on all 176 words as target when preceded each of the 176 words as prime (including itself), measuring the total activation within semantics at every unit update during target presentation. As a proxy for N400 amplitude, we determined the peak in summed activation within the semantic layer after the presentation of each target, and then averaged the sum over a 3-tick window around the peak.[6] The peak always occurred somewhere between 4 and 9 ticks post-onset (out of 16), with a median occurrence at tick 6. Incidentally, assuming that the presentation of each word corresponds to about a second, this range of timing is roughly similar to the actual N400, which is known to occur somewhere between 250 and 500 ms post-stimulus (see Kutas & Federmeier, 2011). Our analyses will consider how N400 amplitude in the model is influenced by the frequency, semantic richness, and orthographic neighborhood size of targets, as well as by whether the prime and target are identical, semantically related (i.e., overlapping in semantic features), associatively related (i.e., the prime-target pair had elevated frequency during training), or unrelated.

2.2. Results and discussion

After training, for 98.7% of target presentations, all semantic units with targets of 1 were more active than all those with targets of 0 on the last tick. All remaining trials involved low-frequency targets and, of these, most involved only one or two incorrect unit activations. Although not perfect, we considered this level of comprehension performance sufficient to warrant testing N400 and behavioral effects in the model.

We will first consider the joint effects of word frequency, semantic richness, and repetition, and then turn to effects of semantic and associative priming and orthographic neighborhood size.

2.2.1. Word frequency, semantic richness, and repetition

The relevant empirical effects on N400 amplitudes are as follows:

Frequency. Low-frequency words produce larger N400s than do high-frequency words (Barber, Vergara, & Carreiras, 2004; Rugg, 1990; Van Petten & Kutas, 1990).

Richness. Words with greater semantic richness—that is, with more semantic features, sometimes operationalized as greater concreteness—yield larger N400s (Kounios & Holcomb, 1994; Kounios et al., 2009; West & Holcomb, 2000) than do words with lower richness.

Repetition. Immediate repetition of a stimulus decreases N400 amplitude (Laszlo & Federmeier, 2007, 2011; Nagy & Rugg, 1989; Sim & Kiefer, 2005).

Frequency × repetition. The effect of repetition in reducing N400 amplitude is greater for low-frequency compared to high-frequency words (Rugg, 1990; Young & Rugg, 1992).

Richness × repetition. The effect of repetition in reducing N400 amplitude is larger for words with greater compared with lesser semantic richness (Rabovsky, Sommer, & Abdel Rahman, 2012; see also Kounios & Holcomb, 1994).

To determine the extent to which the model shows the same pattern of effects, we carried out a three-factor analysis of variance (ANOVA) using the peak amplitude in overall semantic activation, averaged over a 3-tick window (corresponding to the N400 in the model) as the dependent measure. The analysis involved data for each word as target preceded by each word as prime, with target word as the random variable, word frequency and semantic richness as between-item factors, and repetition as a within-item factor.

The pattern of results is shown in Fig. 4. The ANOVA revealed main effects of word frequency ($F_{1,172} = 48.23$, $p < .001$) semantic richness ($F_{1,172} = 336.0$, $p < .001$) and repetition ($F_{1,172} = 862.4$, $p < .001$). In accordance with empirical findings, the simulated N400 was greater for low-frequency words (5.95) compared to high-frequency words (5.15), for high-richness words (6.60) compared to low-richness words (4.50), and for non-repeated words (6.47) compared to repeated words (4.64). In addition, repetition

[5] Although back-propagation is not biologically plausible in literal form, it nonetheless can give rise to internal representations with substantial similarity to neural representations (Khaligh-Razavi & Kriegeskorte, 2014; Kriegeskorte, 2015; Yamins et al., 2014; Zipser & Andersen, 1988), and can be thought of as a computationally efficient approximation of more plausible error-correcting procedures (see O'Reilly, 1996).

[6] All of the simulated N400 results to be reported hold if only the peak itself is used as the dependent measure, but summing over a 3-tick window around the peak provides a more stable measure of the dynamics of semantic activation, and is somewhat more analogous to how empirical data are analyzed. Equivalent results hold if a 5-tick window is used instead.

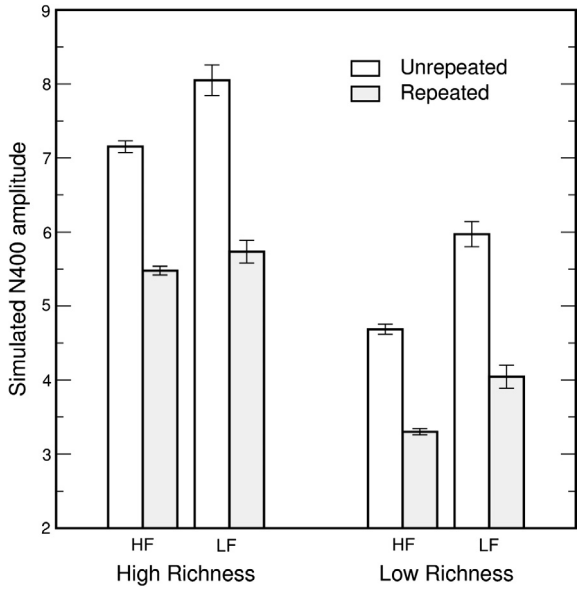

Fig. 4. Means (and standard errors) of simulated N400 amplitudes in the model as a function of word frequency (HF = high-frequency; LF = low-frequency), semantic richness, and repetition.

interacted with both frequency ($F_{1,172} = 22.78,\ p < .001$) and with richness ($F_{1,172} = 7.49,\ p < .01$) such that the reduction due to repetition was greater for low-frequency (2.12) than high-frequency words (1.53) and for high-richness (1.99) than low-richness words (1.65). These interactions are also in agreement with empirical findings. Neither the two-way interaction of frequency and richness ($F_{1,172} = 3.66,\ p = .057$) nor the three-way interaction of frequency, richness and repetition ($F < 1$) were reliable.

In a separate ANOVA using peak time as the dependent measure, there were no reliable effects of frequency, richness, or repetition, nor any interactions.

As many of the relevant empirical studies involved the presentation of words in isolation, rather than in pairs, we also examined the performance of the network on words with no preceding "prime" word. An ANOVA of summed semantic activity (averaged over 3 ticks centered on the peak) with frequency and richness as within-item factors showed reliable effects of both factors (frequency: high 4.70 vs. low 5.66, $F_{1,172} = 59.07,\ p < .001$, (richness: high 6.29 vs. low 4.07, $F_{1,172} = 313.4,\ p < .001$) but no interaction ($F < 1$).

To illustrate that the model, like LP12, produces semantic activation profiles that mirror actual N400 waveforms, as well as to convey a sense of the variability underlying the network's dynamics, Fig. 5 shows the individual activation profiles all words presented in isolation, plotted separately as a function of frequency and richness.

In the model, low-frequency words produce a larger N400 because they are less well learned than high-frequency words and so are less effective at suppressing the features of their orthographic neighbors. High-richness words produce a larger N400 simply because they themselves activate more features than do low-richness words. Repetition reduces the N400 because the target's features suffer decay due to being activated by the prime. This repetition suppression is greater for both low-frequency and high-richness words because the prime-based activation is greater for these two stimulus types. Finally, these variables have little if

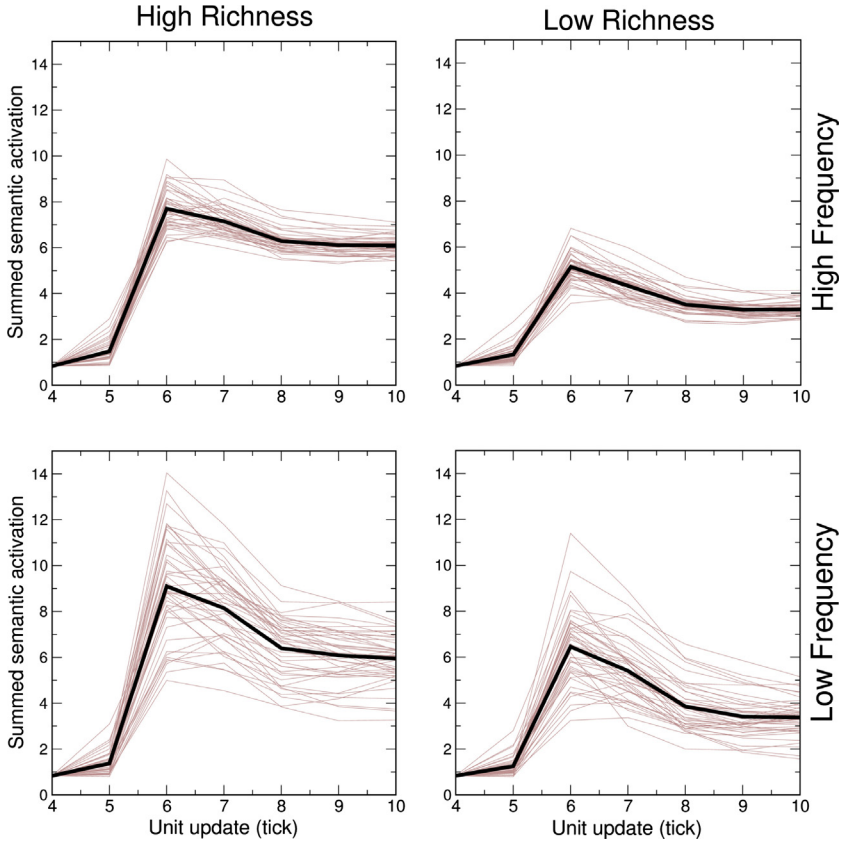

Fig. 5. Summed semantic activation profiles for words varying in frequency and richness when tested in isolation (lighter lines), along with the average of these curves (black line). Unit updates (ticks) are numbered from the onset of the word.

any effect on the latencies of the N400 because the network dynamics depend far more on the architectural organization of excitation and inhibition than on the amount of activation present at any point in time.

2.2.2. Semantic and associative priming

The empirical findings related to semantic priming are made complicated by the fact that different types of relations can fall under the broad notion of "semantic" relatedness (see Moss, Ostrin, Tyler, & Marslen-Wilson, 1995). In particular, researchers have distinguished *associative* relatedness, often measured by free association norms (e.g., DOG-BONE; Postman & Keppel, 1970) from a purely *semantic* relation in which words have similar meanings, such as category coordinates (e.g., DOG-PIG). The problem is that these types of relatedness often co-occur (e.g., DOG-CAT) and, in many studies, stimulus pairs that are characterized as semantically related typically involve both types of relatedness (see Jones, Kintsch, & Mewhort, 2006).

In ERP research, there have been a number of demonstrations that semantic priming decreases N400 amplitude (see Bentin, McCarthy, & Wood, 1985; Federmeier & Kutas, 1999; Kutas, 1993; Kutas & Iragui, 1998; Kutas & Van Petten, 1988) but very few of these studies have attempted to dissociate semantic from associative relatedness. Interestingly, two specific attempts to do so (Koivisto & Revonsuo, 2001; Rhodes & Donaldson, 2008) found clear reductions in N400 amplitude due to associative priming but little if any reduction for pure semantic priming. However, widespread evidence for a modulation of N400 amplitude as a function of congruity of word meanings in context (e.g., "I take coffee with cream and dog/pizza/sugar", Kutas & Hillyard, 1980; see Kutas & Federmeier, 2011 for review) would seem to implicate sensitivity to semantic relatedness as well.

Taken together, then, the empirical evidence from word-word priming paradigms suggests that both associative and semantic relatedness influence N400 amplitudes, with at least some suggestion that the former may be stronger.

The Rabovsky and McRae (2014) model implemented semantic priming in terms of feature overlap between prime and target, but did not address associative priming. The current model includes both associative and semantic relatedness and so their influences can be assessed independently.

Semantic relatedness—semantic feature overlap—was not manipulated as an orthogonal factor in the simulation but varied randomly among word pairs (ranging from 0 to 4 features; with 23.2% of non-identical pairs sharing 1 feature, 2.64% sharing 2 features, and 0.162% sharing 3 or 4 features). Thus, to compare semantic versus associative priming, we calculated mean N400 values for each target when preceded by three types of primes (excluding repetitions): *semantic* primes that were not associatively related but had one or more shared semantic features; *associative* primes that were associates during training but had no semantic feature overlap; and *unrelated* primes that were neither semantically nor associatively related. Forty-seven associates shared one or more semantic features, leaving 129 (pure) associated primes. As shown in Fig. 6, the model shows a small but reliable effect of semantic relatedness on N400 magnitudes (means: 6.38 for semantic primes, 6.50 for unrelated primes; paired $t_{175} = 5.55$, $p < .001$). There is also weak evidence that the degree of relatedness mattered: related pairs with two overlapping feature produced numerically smaller N400s (mean 6.29) than those with only one (mean 6.40), although the difference was only marginally reliable (paired $t_{175} = 1.76$, $p = .08$). The model also showed a clear and somewhat larger effect of associative relatedness (means: 6.09 for associated, 6.35 for unrelated; paired $t_{128} = 3.10$, $p < .005$).

The N400 reduction due to semantic priming is essentially caused by repetition suppression of the shared semantic feature

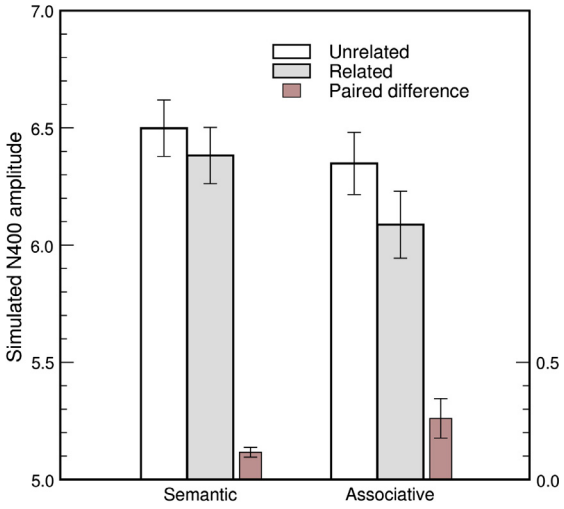

Fig. 6. Mean simulated N400 amplitudes in the model to targets following semantically (but not associatively) related versus unrelated primes, and associatively (but not semantically) related versus unrelated primes. The paired differences between unrelated and related conditions are also plotted (against the right axis). The unrelated conditions differ because they are based on different numbers of observations (176 vs. 129).

(s). Associative priming reduces the N400 because the prime pre-activates its associated target's features to some degree, causing them to suffer from increased decay on presentation of the target. The pre-activation also facilitates learning to suppress the features of the target's orthographic neighbors (much like a high-frequency word).

It should be acknowledged, though, that the relative magnitude of semantic versus associative priming in the model depends on relatively unconstrained properties of the simulation—degree of semantic feature overlap, and prime-target dependencies during training—and, thus, should not be weighted too heavily in evaluating our more general account.

2.2.3. Orthographic neighborhood size

Laszlo and Plaut (2012) showed that simulated N400 amplitudes in their model increased with orthographic neighborhood size for both words and pseudowords, as found empirically (Holcomb, Grainger, & O'Rourke, 2002; Laszlo & Federmeier, 2011). Rabovsky and McRae (2014) observed a small but reliable effect in their model as well. In the current model, as in LP12, there is a small but reliable correlation between orthographic neighborhood size (calculated over words in the training corpus) and N400 amplitude for targets (averaged over all primes; $r = .18$, $t_{174} = 2.37$, $p < .05$). This effect is due to the feedforward excitation coming from the orthographic features that are shared with neighbors; words with more neighbors will thus partially activate the features of a larger number of other words.

2.3. Summary

When a measure of summed semantic activation is used as a proxy for N400 amplitude, the current model exhibits all the relevant empirical effects: a decrease in N400 amplitude for (a) high- vs. low-frequency words; (b) low- vs. high-richness words; (c) repeated vs. non-repeated words; (d) words preceded by semantically or associatively related vs. unrelated primes; and (e) words with smaller vs. larger orthographic neighborhoods, as well as greater repetition effects for (f) low- vs. high-frequency words, and for (g) high- vs. low-richness words. The dynamics of summed

semantic activation also provide a reasonable approximation to the shape of the N400 waveform.

3. Simulation 2: Behavioral effects

Certain issues arise in the current context with regard to modeling behavior. Rabovsky and McRae (2014) used total semantic activation to model behavioral performance in lexical decision, whereas the current model uses this measure to approximate the N400. The problem is that some manipulations, like repetition, actually reduce neural activity while improving performance. Rabovsky and McRae sidestepped this issue by essentially dissociating model activity from neural activity. We, on the other hand, are committed to preserving this relationship.

The mechanism by which reduced neural activity can lead to improved behavioral performance is far from well understood (for discussion, see Gotts, 2015; Gotts, Chow, & Martin, 2012; Henson, Eckstein, Waszak, Frings, & Horner, 2014). A common view is that the overall reduction in neural activity caused by repetition reflects a "sharpening" of neural representations by differentially eliminating the responses of neurons that are relatively poorly-tuned to the stimulus (Desimone, 1996; Wiggs & Martin, 1998). However, careful measurements of neural suppression due to short-term repetitions (on the order of seconds) appear to be more consistent with proportional scaling rather than sharpening (McMahon & Olson, 2007; Miller, Li, & Desimone, 1993; Weiner, Sayres, Vinberg, & Grill-Spector, 2010), and a recent test of this account using an fMRI-adaptation paradigm (Gotts, Milleville, & Martin, 2014) found broadening rather than sharpening of representations following repetition.

An alternative possibility is that the reduction in overall neural activity is accompanied by an increase in spike synchrony between active neurons that make them more effective in driving the downstream neurons responsible for behavior (Ghuman, Bar, Dobbins, & Schnyer, 2008; Gilbert, Gotts, Carver, & Martin, 2010). Although the precise mechanism that gives rise to increased synchrony under repetition has yet to be worked out in detail, there is broad supportive evidence for this account from single-cell recordings (Brunet et al., 2014; Kaliukhovich & Vogels, 2012; Wang, Iliescu, Ma, Josić, & Dragoi, 2011) MEG (Ghuman et al., 2008; Gilbert et al., 2010), and intracranial EEG (Engell & McCarthy, 2014).

Although our computational formalism does not have a means of expressing neural synchrony directly, we formulated a means of generating responses in lexical decision that could take advantage of the information that might drive increased synchrony—namely, the activity-dependent reduction in neural activity. Given that the degree of decay in our model is determined by the time-averaged semantic activations (see Eq. 2), we made these values available to the response system as a proxy for the induced degree of neural synchrony. We recognize that the approximation is not likely to be fully adequate, but considered it the best approach available to us for modeling behavior under repetition suppression.

3.1. Methods

We trained a response network to distinguish words and pseudowords using both time-averaged and instantaneous semantic activation as input. In some ways, the response system can be thought of as a trained approximation of Usher and McClelland's (2001) leaky competing accumulator model of decision making within our neurophysiologically constrained modeling formalism.

3.1.1. Network architecture

We fixed the weights in the comprehension network, and added a new hidden layer of 25 excitatory units and a response (output) layer of 2 excitatory units, corresponding to "yes" and "no" responses. These hidden units received positive-only connections from the semantic units and sent positive-only connections to the response units. The hidden units also had an associated group of 10 "clean-up" units with which it was bidirectionally connected with positive-only weights. The hidden units had an associated inhibitory group of 3 units, and the two response units had a single associated inhibitory unit. Each group sent positive-only connections to, and received negative-only connections from, their corresponding inhibitory group. We also added a copy of the semantic units whose activations were set to the time-averaged activations of the original semantic units (o_j^t in Eq. 2) at every tick. Unlike in the rest of the network, however, the connections from these units to the new hidden layer were not constrained to be positive-only. As discussed earlier, our intent in introducing these units was to make decay-related information available to the response network in whatever way may be useful for improving performance. Although the relationship between the resulting changes in neural synchrony and behavior is not well understood, it is unlikely to reduce to a standard positive-only projection between groups of neurons, and hence there's no reason to constrain the influence of decay-related information in the model in the same way.[7]

Apart from learning rate and momentum, all other parameters were the same as in the comprehension network.

3.1.2. Training procedures

The response network was trained on the 176 words and also on 176 pseudowords that were matched orthographically to the words by selecting randomly from the remaining 324 CVC inputs that were not used as words. Inputs consisted of pairs of stimuli in which a word or pseudoword was followed by a word or pseudoword, where targets were applied only during the last 10 ticks of the second stimulus. The same timing of inputs was used as for the comprehension network—and, for word presentations, the same frequency and associative constraints. The network was trained to activate only the "yes" unit in response to each word, and to activate only the "no" unit in response to each pseudoword.[8] The network was trained for 50,000 presentations of stimulus pairs, using cross-entropy error, a batch size of 1, momentum of 0.8, and a gradually reducing learning rate (0.1 for 25,000 presentations, 0.05 for 25,000 presentations; clean-up layers were again trained with a reduced initial learning rate of 0.001 for the first 5000 presentations).

3.1.3. Testing procedures

We tested the network on all words as targets preceded by all words as primes. As a measure of behavioral performance, we used the sum over the last 6 ticks of the difference between the "yes" and "no" unit activations in response to the target—positive values reflect greater "yes" than "no" activation. We chose this measure because it implicitly reflects both accuracy and latency, as words that activate the "yes" unit and deactivate the "no" unit more quickly and accurately will have higher yes-no sums than words that respond more slowly or less accurately. We used the last 6 ticks because they reflect the steady-state activations reached by the response units after the transient over-activation of the

[7] Indeed, if the outgoing connections from the time-averaged semantic units are constrained to be positive-only, the model shows poorer rather than better performance under repetition, as expected.

[8] We do not, of course, believe that human participants need to be explicitly trained on lexical decision in order to achieve accurate performance on the task, although we do believe that they base their decisions, at least in part, on semantic information (see, e.g., Plaut, 1997). Our use of explicit training on lexical decisions is intended solely to provide a basis for measuring the relative difficulty of saying "yes" to word targets as a function of their properties and relationship to prime words.

N400-like wave (although other numbers of ticks give similar results).

3.2. Results and discussion

3.2.1. Word frequency, semantic richness, and repetition

The relevant empirical effects on behavior are that lexical decision performance is better—in terms of accuracy and/or latency—for high- compared to low-frequency words (Forster & Chambers, 1973; Gardner, Rothkopf, Lapan, & Lafferty, 1987), for high- compared to low-richness words (Pexman, Hargreaves, Siakaluk, Bodner, & Pope, 2008; Yap, Pexman, Wellsby, Hargreaves, & Huff, 2012), and for repeated compared to non-repeated words (Ratcliff, Hockley, & McKoon, 1985; Scarborough, Cortese, & Scarborough, 1977). Importantly, the effects for frequency and repetition are opposite to those for N400 amplitude. The benefit from repetition has been reported to be greater for low- versus high-frequency words (Forster & Davis, 1984; Norris, 1984; although see Versace & Nevers, 2003). Similarly, the repetition benefit has been reported to be greater for high- versus low-richness words (Rabovsky et al., 2012; although see Kounios & Holcomb, 1994).

We carried out a three-factor ANOVA using our behavioral measure (summed yes-no activation) as the dependent measure, target word as the random variable, word frequency and semantic richness as between-item factors, and repetition as a within-item factor (see Fig. 7). The analysis revealed reliable main effects of word frequency ($F_{1,172} = 17.98$, $p < .001$), semantic richness ($F_{1,172} = 6.968$, $p < .01$), and repetition ($F_{1,172} = 28.56$, $p < .001$). As found empirically, the network's performance was better for high-frequency words (4.54) compared to low-frequency words (3.61), for high-richness words (4.37) compared to low-richness words (3.79), and for repeated words (4.37) compared to non-repeated words (3.79). Moreover, repetition interacted with both frequency ($F_{1,172} = 4.573$, $p < .05$) and richness ($F_{1,172} = 30.88$, $p < .001$), such that the repetition benefit was greater for low- than high-frequency words (0.810 vs. 0.347), and greater for high- than low-richness words (1.1803 vs. −0.0231, with the latter not reliably different from 0.0). Frequency and richness did not interact, but the three-way interaction of frequency, richness and repetition was reliable ($F_{1,172} = 13.24$, $p < .001$), because the repetition-by-frequency interaction was much stronger for high- than low-richness words. There is no empirical evidence bearing on the three-way interaction, but the remaining findings are all consistent with those from empirical studies, with the exception of the absence of a repetition effect for low-richness words (cf. Rabovsky et al., 2012).

In the model, words generally activate semantics to a greater degree than pseudowords (after the N400), which means that they suffer from greater decay, as reflected by greater integrated semantic activations in the model. The network thus learns to use these integrated values to support making word responses, and this support is stronger for high- compared to low-frequency words, and for repeated compared to non-repeated words. Semantic activation itself is also informative, which aids high- compared to low-richness words. The interactions arise because the integrated values—which drive the repetition effect—are greater for words with larger N400s: low-frequency words and high-richness words.

3.2.2. Semantic and associative priming

When behavioral studies have tested priming among words that are that are semantically but not associatively related (Fischler, 1977; Moss et al., 1995; Seidenberg, Waters, Sanders, & Langer, 1984; Shelton & Martin, 1992), the priming effect is generally smaller than that found for purely associatively related words, particularly in lexical decision (see McNamara, 2005; Neely, 1991 for reviews; although see Thompson-Schill, Kurtz, & Gabrieli, 1998, for conflicting results).

Using the same analyses as for N400 amplitude but now applied to the behavioral measure (see Fig. 8), the model shows reliable benefits in performance for both semantic priming (means: 3.909 for semantic primes, 3.735 for unrelated primes; paired $t_{175} = 5.505$, $p < .001$) and associative priming (means: 4.314 for associative primes, 3.897 for unrelated primes; paired $t_{128} = 3.704$, $p < .001$), with the latter being larger in magnitude. Thus, in behavior as well as in the N400, the model shows stronger associative than semantic priming, as observed empirically. As was true of the N400 results, semantic priming was numerically greater for primes with two versus one overlapping feature with the target

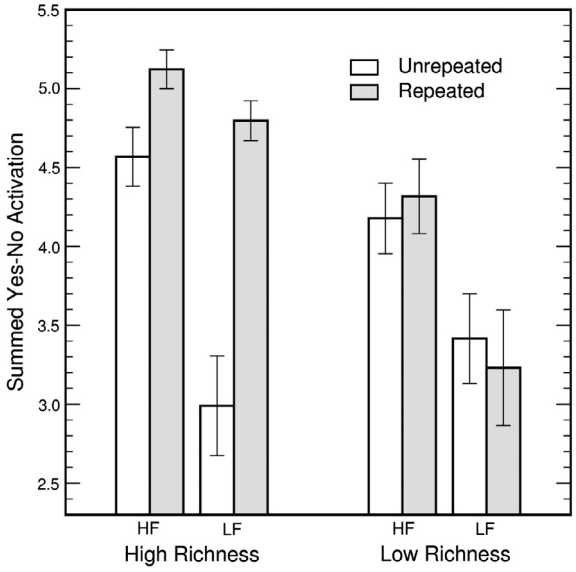

Fig. 7. Means (and standard errors) of the simulated behavioral measure in the model (summed difference in activation between the "yes" and "no" unit in the response system) as a function of word frequency (HF = high-frequency; LF = low-frequency), semantic richness, and repetition.

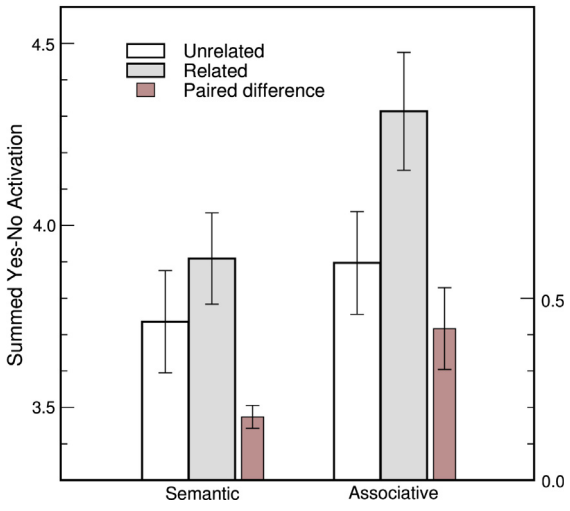

Fig. 8. Mean simulated behavioral measures in the model to targets following semantically (but not associatively) related versus unrelated primes, and associatively (but not semantically) related versus unrelated primes. The paired differences between unrelated and related conditions are also plotted (against the right axis). The unrelated conditions differ because they are based on different numbers of observations (176 vs. 129).

(3.892 vs. 3.735) but the difference was not reliable (paired $t_{175} = 1.356$, $p > .10$).

Both priming effects are driven by greater integrated semantic activations for targets preceded by related compared to unrelated primes.

3.2.3. Orthographic neighborhood size

The effects of orthographic neighborhood size on lexical decision performance are, as Rabovsky and McRae (2014) discuss, rather complicated (see Andrews, 1997; Carreiras, Perea, & Grainger, 1997; Siakaluk, Sears, & Lupker, 2002; Ziegler & Perry, 1998) and depend on factors outside the scope of either model. Interestingly, in the current model the correlation of orthographic neighborhood size and performance is numerically positive and reliable ($r = .29$, $t_{174} = 4.0365$, $p < .001$), due to sensitivity of word responses to greater integrated semantic activations for high-N compared to low-N words.

3.3. Summary

The performance of the response system is broadly successful at modeling the relevant empirical effects. It correctly exhibits better performance for high- versus low-frequency words, high- versus low-richness words, repeated versus non-repeated words, and semantically or associatively primed versus unprimed words, and it also exhibits the empirically observed interactions of repetition with frequency and richness (except that it failed to exhibit a repetition benefit for low-richness words).

4. General discussion

The study of the N400 ERP component, and how it is (or isn't) influenced by various stimulus and context manipulations, has provided a wealth of information on the nature of online comprehension processes (Kutas & Federmeier, 2011), but developing a precise formulation of its mechanistic basis has proved elusive. Laszlo and Plaut (2012; see also Laszlo & Armstrong, 2014) proposed that the N400 corresponds to transient over-activation within semantics due to the distribution of excitation and inhibition found in cortex, and supported their account with a neural network simulation of word comprehension that accounted for effects of orthographic regularity (but not meaningfulness) on single-item N400 amplitudes (Laszlo & Federmeier, 2011).

In contrast to this account, Rabovsky and McRae (2014) proposed that the N400 reflects implicit semantic prediction error, and showed that semantic error in an attractor model of word comprehension (Cree, McRae, & McNorgan, 1999) is influenced in the same manner as is N400 amplitude by a wide range of manipulations, involving word frequency, semantic richness, word repetition, semantic priming, and orthographic neighborhood size. Moreover, semantic activation in their model does a reasonable job of accounting for behavioral effects under the same manipulations. However, in our view, the model and account suffer from a number of shortcomings, the most notable of which are that (a) semantic (prediction) error is not plausibly available under many of the conditions that evoke N400s (e.g., in response to pseudowords); and (b) the dynamics of the error measure over time does not pattern at all similarly to the dynamics of the N400 waveform (see Fig. 2 and compare with Fig. 1a).

In the current work, we apply an extension of the Laszlo and Plaut (2012) model to the same phenomena that Rabovsky and McRae (2014) modeled—adding, among other things, a variant of Laszlo and Armstrong's (2014) activation-based decay function, and an explicit response system. Overall, the current model provides a more satisfactory account of the relevant phenomena. Perhaps most critically, the model—like its predecessors (Laszlo & Armstrong, 2014; Laszlo & Plaut, 2012)—actually produces N400-like deflections in an activation-based measure that could plausibly correspond to the population post-synaptic potentials that underlie ERP components.

It is important to point out that this is not simply a case in which there are two models that both account for the same set of findings—the nature of what it means to "account" for a finding is fundamentally different on the two approaches. Rabovsky and McRae (2014) identified variables in their model—error and summed semantic activation—that are influenced in the same directions as empirically observed effects on N400 amplitude and behavioral performance, respectively. However, doing so does not provide a *mechanistic* account of the empirical phenomena—not withstanding the use of computational modeling—unless the model variables can be linked to participants' neural and cognitive mechanisms in a plausible manner. The fact that semantic error is based on information that is unavailable to participants, and does not exhibit the signature temporal dynamics of the physiological measure to which it putatively corresponds, undermines for us the relevance of the model to understanding comprehension mechanisms in brain and behavior. By contrast, the current model—despite its many limitations (as discussed below)—provides a more direct and therefore more informative mapping between real-time activation processes in the model and real-time activation processes in the brain. In this way, the approach offers the beginning of an *explanation* of the relevant empirical phenomena.

4.1. Relation to Rabovsky and McRae (2014) account

To be clear, we think there is much to recommend Rabovsky and McRae's (2014) theoretical emphasis on prediction (see also Kuperberg & Jaeger, 2015), and, in fact, it aligns with our own perspective under many conditions. Consider associative priming. In terms of prediction error, the prime leads to an elevated expectation of the occurrence of the target's features, and thus when the target actually occurs there is less prediction error than when the target is unexpected following an unrelated prime. Note, though, that the same thing is true in terms of degree of over-activation of semantics (under the proper constraints on excitation and inhibition): the prime partially pre-activates features of the associated target, giving those features an advantage in—and, thus, shortening—the subsequent competition when the target presentation activates features of its orthographic neighbors. On this latter account, any source of pre-activation of appropriate semantic features, including sentence-level and discourse-level context, would be expected to reduce N400 amplitudes. Indeed, the word-level effect of semantic/associative priming on the N400 is indistinguishable from the sentence-level effect on final words of congruent versus incongruent sentences (Kutas, 1993). Thus, our account and one based on prediction error agree in cases where any kind of prior context pre-activates (or "predicts") semantic features. We prefer our account account in part because it maintains a clear relationship between simulated neural activity and the EEG signal. Neurally explicit formulations of predictive coding (e.g., Friston, 2010; Park & Friston, 2013) typically employ a population of "prediction error" neurons that are separate from more conventional "representational" neurons, but it is unclear why, on the RM14 account, the EEG signal would reflect only the former (although see Friston, 2005).

Another advantage of an account based on transient over-activation is that it also applies in the absence of informative context and, hence, in the absence of a basis for making predictions. Consider the word frequency effect. In our model, low-frequency words generate larger N400s because they have not learned to

suppress their neighbors' features as well as have high-frequency words. According to Rabovsky and McRae (2014), implicit prediction error can explain the word frequency effect because:

> An internal model should encode the fact that, in general, it is less probable to encounter a low frequency as compared to a high frequency word. Therefore, implicit prediction error would be higher for low frequency words. (p. 71)

This sounds plausible but works only if one assumes localist word representations, so that the units for high-frequency words can be pre-activated more than those for low-frequency words. It doesn't work for the distributed (semantic) representations in their model and ours: the *features* of high-frequency words are no more common (and hence would not be more strongly pre-activated) than those of low-frequency words. Thus, the word MILK is very high-frequency but "produced by cows" is certainly not; AQUIFER is very low-frequency but being "related to water" is relatively common.

The Rabovsky and McRae (2014) model produces lower prediction error for high-frequency words because the stronger activations for such words are compared against correct semantic targets, even though these are unavailable from implicit prediction. Stronger semantic activation might very well produce higher error if compared against the type of generic predictions that are actually possible in random word lists.

This issue is, of course, even starker in the case of nonwords, including illegal strings and pseudowords, which don't have correct semantic values to compute error against. Rabovsky and McRae (2014) did not apply their model to any nonword stimuli, but suggested (p. 84) that "illegal strings presumably correctly elicit very little expectation for semantic features at all, so that implicit prediction error would be low". The problem with this suggestion is that prediction error is low only if the presumed low levels of semantic activation are compared against "correct" semantic targets of all zeros, and yet the system has no way to know that these are the correct targets until *after* the stimulus has been processed and determined to be a nonword. The N400 itself reflects this processing, and thus an account of the N400 cannot presume it has already occurred.

Repetition effects are perhaps the clearest example of reduction in N400 amplitude due to pre-activating semantic features. On our account, though, this alone is insufficient to give rise to the pattern of interactions of repetition with semantic richness and, in particular, with word frequency. In both cases, reductions are greatest for the items that produce the largest initial N400—high-richness words and low-frequency words. But in the latter case, the source of the larger N400 is the activation of neighbors' features, not those of the word itself. Pre-activation of competitors' features gives no advantage to the features of the low-frequency word on the second presentation (to resolve the competition more quickly)—quite the opposite in fact. Rather, the interactions with repetition, and much of the main effect of repetition itself, are due to the operation of activation-dependent decay. Greater overall activation during the first presentation leads to greater decay on those active features, and thus a larger reduction in N400 on the second presentation.

Laszlo and Armstrong (2014) introduced the idea that activity-dependent decay formed the basis for reductions in N400 amplitudes following repetition. They employed an *alpha* function which is used to model fatigue effects in post-synaptic potentials (Bugmann, 1997), which underlie ERP signals (Fabiani et al., 2007), and related functions have been shown to approximate activation dynamics in actual neurons (David et al., 2006; see also Gotts & Plaut, 2002, for related modeling of the relevance of synaptic depression to comprehension impairments). We chose to adopt a version of activity-dependent decay that is somewhat simplified relative to the alpha function but gives rise to qualitatively similar effects. Although the specific decay function may not matter much, we do think that some form of repetition suppression is critical to accounting for repetition effects on the N400, and context effects more generally.

Following Rabovsky and McRae (2014), we attempted to model not just the electrophysiological consequences of the various factors but also their impact on behavior. Rabovsky and McRae assumed that greater semantic activation corresponds to better performance, and showed that this largely aligned with human performance for their model (although sometimes very weakly, and not always over the time range corresponding to response generation). However, this approach in untenable in our model—and, we believe, in any model that incorporates repetition suppression—because some conditions that give rise to better performance, such as repetition, actually produce weaker overall activation. As discussed earlier, there is as yet no clear explanation for improved performance under repetition suppression (see Gotts et al., 2012), but one promising possibility is that the reduced neural activation increases neural synchrony which, in turn, increases the efficacy of downstream communication (Ghuman et al., 2008; Gilbert et al., 2010).

Given these considerations, we decided against stipulating a particular measure as corresponding to behavioral performance, but rather provided a response system with potentially relevant information and allowed it to learn to produce accurate behavior. As our framework does not have a means of expressing neural synchrony directly, we provided the response system with information that is thought to govern synchrony—the degree of activation-based decay, as determined by each semantic unit's time-averaged activation.

The performance of the resulting trained response system does accord with the observed empirical effects (apart from the repetition benefit for low-richness words; Rabovsky et al., 2012). Even so, further work is needed to replicate these findings using better approximations to the effects of repetition on neural synchrony, and of neural synchrony on response generation.

4.2. Limitations and future directions

There is no question that our model of word comprehension suffers from a number of limitations in its design and scope, and understanding these is critical to informing the development of better models in the future. In addition to the issues related to neural synchrony just mentioned, the small size of the vocabulary, the artificiality of the semantic representations, and the implausibility of the learning procedure are all aspects that could be improved. But perhaps more fundamental than these is the restriction to sequences of pairs of words, and to deriving isolated word meanings rather than sentence- or discourse-level interpretations. A large proportion of the literature on the N400 concerns its sensitivity, or lack thereof, to sentence-level contexts and manipulations. For this reason, a critical extension of the current work would be to apply the same computational principles and account within a model of sentence comprehension (e.g., McClelland, St. John, & Taraban, 1989; St. John & McClelland, 1990). It is also important to extend the approach to address the properties of other comprehension-related ERP components, such as the P600 and its apparent complementary sensitivity to syntactic but not semantic violations (Friederici, 1995; Hagoort, Brown, & Groothusen, 1993; but see Brouwer & Hoeks, 2013; Brouwer, Fitz, & Hoeks, 2012; Brouwer, Crocker, Venhuizen, & Hoeks, accepted for publication; Kuperberg, 2007 for an interesting alternative perspective in which the relevant distinction is between lexical-level vs. sentence-level integration).

It is also important to acknowledge that the current treatment of the relationship between model activity, neural activity, and

the EEG signal is highly simplified and in need of elaboration. We make the standard assumption that the real-valued sigmoid activation function approximates neural firing frequency relative to some maximal rate (Cohen & Servan-Schreiber, 1992), and our activation-based decay function can be interpreted as approximating neural adaptation due to synaptic depression (Abbott, Varela, Sen, & Nelson, 1997; Gotts & Plaut, 2002; Varela et al., 1997). We also assume that summed activation within a layer of the model is a sufficient approximation of the population-based post-synaptic potentials underlying EEG signals (Fabiani et al., 2007). However, these assumptions are clearly inadequate in light of a consideration of neural oscillations. First, the efficacy of neural communication is not solely a function of firing rate but also of the degree of synchrony among incoming action potentials (see Konig, Engel, & Singer, 1996; Salinas & Sejnowski, 2001; Singer, 1999), and our introduction of integrated semantic activity is, at best, a poor approximation to this. Moreover, a number of researchers have argued that neural oscillations are directly relevant to interpreting ERP components like the N400 (see Makeig et al., 2002; Roehm, Schlesewsky, Bornkessel, Frisch, & Haider, 2004; Sauseng et al., 2007). Nonetheless, we believe it is prudent to explore and understand the limitations of simpler accounts (i.e., overall neural activity) before introducing more complexity. We see no fundamental problem with extending a model based on neural activity to include a consideration of neural oscillations and synchrony.

4.3. Conclusions

We have presented an extension of computational work by Laszlo and colleagues (Laszlo & Armstrong, 2014; Laszlo & Plaut, 2012) in which the N400 ERP component corresponds to transient over-activation within semantics, due to the intrinsic distribution of excitation and inhibition within and between cortical areas. The model accounts for the same range of ERP and behavioral findings as an alternative model based on semantic prediction error (Rabovsky & McRae, 2014). The two accounts broadly agree on the basis for the effects of prior context on the N400. However, we believe that the current account has a number of important advantages, including the fact that it actually produces N400 morphology, is based solely on neural activation rather than implausible access to correct semantic information, and can explain N400 effects even for meaningless stimuli (e.g., pseudowords). Although considerable work remains in improving the scale of the simulation and in applying the approach to a broader range of phenomena, including sentence-level effects, the current findings further contribute to establishing the value of developing computationally explicit theories of the relationship between brain function and behavior.

Acknowledgment

This work was supported by a year-long Undergraduate Research Fellowship in Computational Neuroscience at Carnegie Mellon University to SJC, under NIH Grant R90DA023426. We thank the VisCog research group at Carnegie Mellon University for helpful discussions, and Blair Armstrong, Marlene Behrmann, and Sarah Laszlo for detailed comments on the paper.

References

Abbott, L. F., Varela, K., Sen, K., & Nelson, S. B. (1997). Synaptic depression and cortical gain control. *Science, 275*, 220–223.

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review, 4*, 439–461.

Barber, H., Vergara, M., & Carreiras, M. (2004). Syllable-frequency effects in visual word recognition: Evidence from ERPs. *Neuroreport, 15*, 545–548. http://dx.doi.org/10.1097/01.wnr.0000111325.38420.80.

Bentin, S., McCarthy, G., & Wood, C. C. (1985). Event-related potentials, lexical decision and semantic priming. *Electroencephalography and Clinical Neurophysiology, 60*, 343–355.

van Berkum, J. J. A. (2009). The neuropragmatics of "simple" utterance comprehension: An ERP review. In U. Sauerland & K. Yatsushiro (Eds.), *Semantics and pragmatics: From experiment to theory* (pp. 276–316). Basingstoke, UK: Palgrave Macmillan.

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. J. (accepted for publication). A neurocomputational model of the N400 and the P600 in language processing. *Cognitive Science*.

Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about semantic illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research, 1446*, 127–143. http://dx.doi.org/10.1016/j.brainres.2012.01.055.

Brouwer, H., & Hoeks, J. C. J. (2013). A time and place for language comprehension: Mapping the N400 and the P600 to a minimal cortical network. *Frontiers in Human Neuroscience, 7*, 758. http://dx.doi.org/10.3389/fnhum.2013.00758.

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

Brunet, N. M., Bosman, C. A., Vinck, M., Roberts, M., Oostenveld, R., Desimone, R., ... Fries, P. (2014). Stimulus repetition modulates gamma-band synchronization in primate visual cortex. *Proceedings of the National Academy of Science, USA, 111*, 3626–3631.

Bugmann, G. (1997). Biologically plausible neural computation. *Biosystems, 40*, 11–19.

Carreiras, M., Perea, M., & Grainger, J. (1997). Effects of orthographic neighborhood in visual word recognition: Cross-task comparisons. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 857–871.

Cohen, J. D., & Servan-Schreiber, D. (1992). Context, cortex, and dopamine: A connectionist approach to behavior and biology in schizophrenia. *Psychological Review, 99*, 45–77.

Cree, G. S., McNorgan, C., & McRae, K. (2006). Distinctive features hold a privileged status in the computation of word meaning: Implications for theories of semantic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 643–658. http://dx.doi.org/10.1037/0278-7393.32.4.643.

Cree, G. S., McRae, K., & McNorgan, C. (1999). An attractor model of lexical conceptual processing: Simulating semantic priming. *Cognitive Science, 23*, 371–414.

Crick, F. H. C. (1989). The recent excitement about neural networks. *Nature, 337*, 129–132.

David, O., Kiebel, S. J., Harrison, L. M., Mattout, J., Kilner, J. M., & Friston, K. J. (2006). Dynamic causal modeling of evoked responses in EEG and MEG. *Neuroimage, 30*, 1255–1272.

Deacon, D., Dynowska, A., Ritter, W., & Grose-Fifer, J. (2004). Repetition and semantic priming of nonwords: Implications for theories of N400 and word recognition. *Psychophysiology, 41*, 60–74.

Deacon, D., Dynowska, A., Ritter, W., & Grose-Fifver, J. (2009). Repetition and semantic priming of nonwords: Implications for theories of N400 and word recognition. *Psychophysiology, 41*, 60–74.

Desimone, R. (1996). Neural mechanisms for visual memory and their role in attention. *Proceedings of the National Academy of Science, USA, 93*, 13494–13499.

Engell, A. D., & McCarthy, G. (2014). Repetition suppression of face-selective evoked and induced EEG recorded from the human cortex. *Human Brain Mapping, 35*, 4155–4162.

Fabiani, M., Gratton, G., & Federmeier, K. D. (2007). Event-related brain potentials: Methods, theory, and application. In J. T. Cacioppo, L. Tassinary, & G. Berntson (Eds.), *Handbook of psychophysiology* (3rd ed., pp. 85–119). Cambridge, UK: Cambridge University Press.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language, 41*, 469–495.

Federmeier, K. D., & Laszlo, S. (2009). Time for meaning: Electrophysiology provides insights into the dynamics of representation and processing in semantic memory. In B. Ross (Ed.), *Psychology of learning and memory* (pp. 1–44). Burlington, MA: Academic Press.

Fischler, I. (1977). Semantic facilitation without association in a lexical decision task. *Memory and Cognition, 5*, 335–339.

Forster, K. I., & Chambers, S. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behaviour, 12*, 627–635.

Forster, K. I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 680–698.

Friederici, A. D. (1995). The time course of syntactic activation during language processing: A model based on neuropsychological and neurophysiological data. *Brain and Language, 50*, 259–281.

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London, Series B, 360*, 815–836. http://dx.doi.org/10.1098/rstb.2005.1622.

Friston, K. J. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*, 127–138. http://dx.doi.org/10.1038/nrn2787.

Gardner, M. K., Rothkopf, E. Z., Lapan, R., & Lafferty, T. (1987). The word frequency effect in lexical decision: Finding a frequency-based component. *Memory and Cognition, 15*, 24–28.

Ghuman, A. S., Bar, M., Dobbins, I. G., & Schnyer, D. M. (2008). The effects of priming on frontal-temporal communication. *Proceedings of the National Academy of Science, USA, 105*, 8405–8409.

Gilbert, J. R., Gotts, S. J., Carver, F. W., & Martin, A. (2010). Object repetition leads to local increases in the temporal coordination of neural responses. *Frontiers in Human Neuroscience, 4*, 30. http://dx.doi.org/10.3389/fnhum.2010.00030.

Gotts, S. J. (2015). Incremental learning of perceptual and conceptual representations and the puzzle of neural repetition suppression. *Psychonomic Bulletin & Review, 23*, 1055–1071. http://dx.doi.org/10.3758/s13423-015-0855-y.

Gotts, S. J., Chow, C. C., & Martin, A. (2012). Repetition priming and repetition suppression: Multiple mechanisms in need of testing. *Cognitive Neuropsychology, 3*, 250–259. http://dx.doi.org/10.1080/17588928.2012.697054.

Gotts, S. J., Milleville, S. C., & Martin, A. (2014). Object identification leads to a conceptual broadening of object representations in lateral prefrontal cortex. *Neuropsychologia, 76*, 62–78. http://dx.doi.org/10.1016/j.neuropsychologia.2014.10.041.

Gotts, S. J., & Plaut, D. C. (2002). The impact of synaptic depression following brain damage: A connectionist account of "access/refractory" and "degraded-store" semantic impairments. *Cognitive, Affective and Behavioral Neuroscience, 2*, 187–213.

Hagoort, P., Baggio, G., & Willems, R. M. (2009). Semantic unification. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 819–836). Boston, MA: MIT Press.

Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and Cognitive Processes, 8*, 439–483. http://dx.doi.org/10.1080/01690969308407585.

Halgren, E., Dhond, R. P., Christensen, N., Van Petten, C., Marinkovic, K., Lewine, J. D., et al. (2002). N400-like magnetoencephalography responses modulated by semantic context, word frequency, and lexical class in sentences. *Neuroimage, 17*, 1101–1116. http://dx.doi.org/10.1006/nimg.2002.1268.

Henson, R. N., Eckstein, D., Waszak, F., Frings, C., & Horner, A. J. (2014). Stimulus-response bindings in priming. *Trends in Cognitive Sciences, 18*, 376–384. http://dx.doi.org/10.1016/j.tics.2014.03.004.

Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review, 98*, 74–95.

Holcomb, P. J., Grainger, J., & O'Rourke, T. (2002). An electrophysiological study of the effects of orthographic neighborhood size on printed word perception. *Journal of Cognitive Neuroscience, 14*, 938–950.

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language, 55*, 534–552.

Kaliukhovich, D. A., & Vogels, R. (2012). Stimulus repetition affects both strength and synchrony of macaque inferior temporal cortical activity. *Journal of Neurophysiology, 107*, 3509–3527.

Kandel, E. R., & Schwartz, J. H. (1985). *Principles of neural science*. New York: Elsevier.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology, 10*, e1003915. http://dx.doi.org/10.1371/journal.pcbi.1003915.

Koivisto, M., & Revonsuo, A. (2001). Cognitive representations underling the N400 priming effect. *Cognitive Brain Research, 12*, 467–490.

Konig, P., Engel, A. K., & Singer, W. (1996). Integrator or coincidence detector? The role of the cortical neuron revisited. *Trends in Neurosciences, 19*, 202–208.

Kounios, J., Green, D. L., Payne, L., Fleck, J. I., Grondin, R., & McRae, K. (2009). Semantic richness and the activation of concepts in semantic memory: Evidence from event-related potentials. *Brain Research, 1282*, 95–102. http://dx.doi.org/10.1016/j.brainres.2009.05.092.

Kounios, J., & Holcomb, P. J. (1994). Concreteness effects in semantic processing: ERP evidence supporting dual-coding theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 804–823.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science, 1*, 417–446. http://dx.doi.org/10.1146/annurev-vision-082114-035447.

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

Kuperberg, G. R., & Jaeger, T. F. (2015). What do we mean by prediction in language comprehension? *Language, Cognition & Neuroscience, 31*, 32–59. http://dx.doi.org/10.1080/23273798.2015.1102299.

Kutas, M. (1993). In the company of other words: Electrophysiological evidence for single-word and sentence context effects. *Language and Cognitive Processes, 8*, 533–572.

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. (2009). N400. *Scholarpedia, 4*, 7790. http://dx.doi.org/10.4249/scholarpedia.7790.

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. http://dx.doi.org/10.1146/annurev.psych.093008.131123.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207*, 203–205.

Kutas, M., & Iragui, V. (1998). The N400 in a semantic categorization task across 6 decades. *Electroencephalography and Clinical Neurophysiology, 108*, 456–471.

Kutas, M., & Van Petten, C. (1988). Event-related brain potential studies of language. In P. K. Ackles, J. R. Jennings, & M. G. H. Coles (Eds.), *Advances in psychophysiology* (pp. 139–187). Greenwich, CT: JAI Press.

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.

Laszlo, S., & Federmeier, K. D. (2007). Better the DVL you know: Acronyms reveal the contribution of familiarity to single word reading. *Psychological Science, 18*, 122–126.

Laszlo, S., & Federmeier, K. D. (2009). A beautiful day in the neighborhood: An event-related potential study of lexical relationships and prediction in context. *Journal of Memory and Language, 61*, 326–338.

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.

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.

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: [De]constructing the N400. *Nature Reviews Neuroscience, 9*, 920–933.

Makeig, S., Westerfield, M., Jung, T. P., Enghoff, S., Townsend, J., Courchesne, E., et al. (2002). Dynamic brain sources of visual evoked responses. *Science, 295*, 690–694.

Marr, D. (1971). Simple memory: A theory for archicortex. *Proceedings of the Royal Society of London, Series B, 262*, 23–81. http://dx.doi.org/10.1098/rstb.1971.0078.

McClelland, J. L., St. John, M., & Taraban, R. (1989). Sentence comprehension: A parallel distributed processing approach. *Language and Cognitive Processes, 4*, 287–335.

McMahon, D. B., & Olson, C. R. (2007). Repetition suppression in monkey inferotemporal cortex: Relation to behavioral priming. *Journal of Physiology, 97*, 3532–3543.

McNamara, T. P. (2005). *Semantic priming: Perspectives from memory and word recognition*. New York: Psychology Press.

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

Miller, E. K., Gochin, P. M., & Gross, C. G. (1991). Habituation-like decrease in the responses of neurons in inferior temporal cortex of the macaque. *Visual Neuroscience, 7*, 357–362.

Miller, E. K., Li, L., & Desimone, R. (1993). Activity of neurons in anterior inferior temporal cortex during a short-term memory task. *Journal of Neuroscience, 13*, 1460–1478.

Moss, H. E., Ostrin, R. K., Tyler, L. K., & Marslen-Wilson, W. D. (1995). Accessing different types of lexical semantic information: Evidence from priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 863–883.

Nagy, M. E., & Rugg, M. D. (1989). Modulation of event-related potentials by word repetition: The effects of inter-item lag. *Psychophysiology, 26*, 431–436.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading* (pp. 264–336). Hillsdale, NJ: Lawrence Erlbaum Associates.

Norris, D. (1984). The effects of frequency, repetition and stimulus quality in visual word recognition. *Quarterly Journal of Experimental Psychology, Section A: Human Experimental Psychology, 36*, 507–518.

O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation, 8*, 895–938.

Park, H.-J., & Friston, K. (2013). Structural and functional brain networks: From connections to cognition. *Science, 342*, 1238411. http://dx.doi.org/10.1126/science.1238411.

Pearlmutter, B. A. (1989). Learning state space trajectories in recurrent neural networks. *Neural Computation, 1*, 263–269.

Pexman, P. M., Hargreaves, I. S., Siakaluk, P. D., Bodner, G. E., & Pope, J. (2008). There are many ways to be rich: Effects of three measures of semantic richness on visual word recognition. *Psychonomic Bulletin & Review, 15*, 161–167. http://dx.doi.org/10.3758/Pbr.15.1.161.

Plaut, D. C. (1995). Semantic and associative priming in a distributed attractor network. In *Proceedings of the 17th annual conference of the Cognitive Science Society* (pp. 37–42). Hillsdale, NJ: Lawrence Erlbaum Associates.

Plaut, D. C. (1997). Structure and function in the lexical system: Insights from distributed models of naming and lexical decision. *Language and Cognitive Processes, 12*, 767–808.

Plaut, D. C., & Booth, J. R. (2000). Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review, 107*, 786–823.

Postman, L., & Keppel, G. (1970). *Norms of word associations*. New York: Academic Press.

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*, 68–98. http://dx.doi.org/10.1016/j.cognition.2014.03.010.

Rabovsky, M., Sommer, W., & Abdel Rahman, R. (2012). Implicit word learning benefits from semantic richness: Electrophysiological and behavioral evidence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 1076–1083. http://dx.doi.org/10.1037/a0025646.

Ratcliff, R., Hockley, W., & McKoon, G. (1985). Components of activation: Repetition and priming effects in lexical decision and recognition. *Journal of Experimental Psychology: General, 114*, 435–450.

Rhodes, S. M., & Donaldson, D. I. (2008). Association and not semantic relationships elicit the N400 effect: Electrophysiological evidence from an explicit language comprehension task. *Psychophysiology, 45*, 50–59. http://dx.doi.org/10.1111/j.1469-8986.2007.00598.x.

Roehm, D., Schlesewsky, M., Bornkessel, I., Frisch, S., & Haider, H. (2004). Fractionating language comprehension via frequency characteristics of the human EEG. *Neuroreport, 15*, 409–412.

Rugg, M. D. (1990). Event-related brain potentials dissociate repetition effects of high-frequency and low-frequency words. *Memory and Cognition, 18*, 367–379.

Salinas, E., & Sejnowski, T. J. (2001). Correlated neuronal activity and the flow of neural information. *Nature Neuroscience Reviews, 2*, 539–550.

Sauseng, P., Klimesch, W., Gruber, W. R., Hanslmayr, S., Freunberger, R., & Doppelmayr, M. (2007). Are event-related potential components generated by phase resetting of brain oscillations? A critical discussion. *Neuroscience, 146*, 1435–1444. http://dx.doi.org/10.1016/j.neuroscience.2007.03.014.

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance, 3*, 1–17.

Seidenberg, M. S., Waters, G. S., Sanders, M., & Langer, P. (1984). Pre- and postlexical loci of contextual effects on word recognition. *Memory and Cognition, 12*, 315–328.

Shelton, J. R., & Martin, R. C. (1992). How semantic is automatic semantic priming? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 1191–1210.

Siakaluk, P. D., Sears, C. R., & Lupker, S. J. (2002). Orthographic neighborhood effects in lexical decision: The effects of nonword orthographic neighborhood size. *Journal of Experimental Psychology: Human Perception and Performance, 28*, 661–681.

Sim, E. J., & Kiefer, M. (2005). Category-related brain activity to natural categories is associated with the retrieval of visual features: Evidence from repetition effects during visual and functional judgments. *Cognitive Brain Research, 24*, 260–273. http://dx.doi.org/10.1016/j.cogbrainres.2005.02.006.

Singer, W. (1999). Striving for coherence. *Nature, 397*, 391–393.

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

Thompson-Schill, S. L., Kurtz, K. J., & Gabrieli, J. D. E. (1998). Effects of semantic and associative relatedness on automatic priming. *Journal of Memory and Language, 38*, 440–458.

Usher, M., & McClelland, J. L. (2001). On the time course of perceptual choice: The leaky competing accumulator model. *Psychological Review, 108*, 550–592.

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

Van Petten, C., & Luka, B. J. (2006). Neural localization of semantic context effects in electromagnetic and hemodynamic studies. *Brain and Language, 97*, 279–293.

Varela, J., Sen, K., Gibson, J., Fost, J., Abbott, L. F., & Nelson, S. B. (1997). A quantitative description of short-term plasticity at excitatory synapses in layer 2/3 of rat primary visual cortex. *Journal of Neurophysiology, 17*, 7926–7940.

Versace, R., & Nevers, B. (2003). Word frequency effect on repetition priming as a function of prime duration and delay between the prime and the target. *British Journal of Psychology, 94*, 389–480.

Wang, Y., Iliescu, B. F., Ma, J., Josić, K., & Dragoi, V. (2011). Adaptive changes in neuronal synchronization in macaque V4. *Journal of Neuroscience, 31*, 13204–13213.

Weiner, K. S., Sayres, R., Vinberg, J., & Grill-Spector, K. (2010). fMRI-adaptation and category selectivity in human ventral temporal cortex: Regional differences across time scales. *Journal of Neurophysiology, 103*, 3349–3356.

West, W. C., & Holcomb, P. J. (2000). Imaginal, semantic, and surface-level processing of concrete and abstract words: An electrophysiological investigation. *Journal of Cognitive Neuroscience, 12*, 1024–1037.

Wiggs, C. L., & Martin, A. (1998). Properties and mechanisms of perceptual priming. *Current Opinion in Neurobiology, 8*, 227–233.

Yamins, D. L. K., Honga, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Science, USA, 111*, 8619–8624. http://dx.doi.org/10.1073/pnas.1403112111.

Yap, M. J., Pexman, P. M., Wellsby, M., Hargreaves, I. S., & Huff, M. J. (2012). An abundance of riches: Cross-task comparisons of semantic richness effects in visual word recognition. *Frontiers in Human Neuroscience, 6*, 72. http://dx.doi.org/10.3389/fnhum.2012.00072.

Young, M. P., & Rugg, M. D. (1992). Word-frequency and multiple repetition as determinants of the modulation of event-related potentials in a semantic classification task. *Psychophysiology, 29*, 664–676.

Zheng, Y., Luo, J. J., Harris, S., Kennerley, A., Berwick, J., Billings, S. A., et al. (2012). Balanced excitation and inhibition: Model based analysis of local field potentials. *NeuroImage, 63*, 81–94. http://dx.doi.org/10.1016/j.neuroimage.2012.06.040.

Ziegler, J. C., & Perry, C. (1998). No more problems in Coltheart's neighborhood: Resolving neighborhood conflicts in the lexical decision task. *Cognition, 68*, B53–B62.

Zipser, D., & Andersen, R. A. (1988). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature, 331*, 679–684.