Cognition 132 (2014) 68–89

Contents lists available at ScienceDirect

Cognition

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

Simulating the N400 ERP component as semantic network error: Insights from a feature-based connectionist attractor model of word meaning

Milena Rabovsky [a,*], Ken McRae [b]

[a] Department of Psychology, Humboldt-Universität zu Berlin, Germany
[b] Department of Psychology, University of Western Ontario, London, Canada

ARTICLE INFO

Article history:
Received 10 November 2012
Revised 12 March 2014
Accepted 28 March 2014
Available online 22 April 2014

Keywords:
N400
ERP
Connectionism
Network model
Prediction error
Semantic system

ABSTRACT

The N400 ERP component is widely used in research on language and semantic memory. Although the component's relation to semantic processing is well-established, the computational mechanisms underlying N400 generation are currently unclear (Kutas & Federmeier, 2011). We explored the mechanisms underlying the N400 by examining how a connectionist model's performance measures covary with N400 amplitudes. We simulated seven N400 effects obtained in human empirical research. Network error was consistently in the same direction as N400 amplitudes, namely larger for low frequency words, larger for words with many features, larger for words with many orthographic neighbors, and smaller for semantically related target words as well as repeated words. Furthermore, the repetition-induced decrease was stronger for low frequency words, and for words with many semantic features. In contrast, semantic activation corresponded less well with the N400. Our results suggest an interesting relation between N400 amplitudes and semantic network error. In psychological terms, error values in connectionist models have been conceptualized as implicit prediction error, and we interpret our results as support for the idea that N400 amplitudes reflect implicit prediction error in semantic memory (McClelland, 1994).

1. Introduction

The N400 component of the event-related brain potential (ERP) is used widely in neuro-cognitive research on language and semantic memory. Although the component's relation to semantic processing is well-established, the specific computational mechanisms underlying N400 generation are currently not clear (Kutas & Federmeier, 2011). The goal of the present research is to elucidate this issue by relating N400 amplitude modulations to variations in performance measures of an attractor network model of word meaning.

In the remainder of the Introduction, we first present a short review of empirical N400 data. We then discuss theories of the cognitive sources of the N400, as well as Laszlo and Plaut's (2012) connectionist simulations of the N400. Next, we outline our approach to modeling the N400, which differs in key ways from that of Laszlo and Plaut. In this discussion, we define "implicit prediction error" as it is used in the present article and simulations.

1.1. N400 data

The N400 is a negative-going ERP component with a broad centro-parietal scalp distribution peaking at about

[*] Corresponding author. Address: Department of Psychology, Humboldt-Universität zu Berlin, Rudower Chaussee 18, 12489 Berlin, Germany. Tel.: +49 30 2093 4904; fax: +49 30 2093 9332.

E-mail address: milena.rabovsky@hu-berlin.de (M. Rabovsky).

http://dx.doi.org/10.1016/j.cognition.2014.03.010

400 ms after stimulus presentation, and is linked to the processing of meaning (see Kutas & Federmeier, 2011, for comprehensive review). It was initially reported by Kutas and Hillyard (1980), who showed that N400 amplitudes were modulated by a word's fit in a sentence context, with much larger amplitudes for contextually anomalous than for predictable words. For example, the N400 is larger when "I take my coffee with cream and ..." is completed by "socks" as compared to "sugar". Subsequent work showed that N400 amplitudes gradually decrease with increasing expectancy of a given word in a given context (Kutas & Hillyard, 1984). However, a sentence context is not needed to elicit N400 modulations. For example, N400 amplitudes vary as a function of semantic relatedness between word pairs, with smaller amplitudes to a target word following a related (*van–truck*) as compared to an unrelated (*hat–truck*) prime (Bentin, Mccarthy, & Wood, 1985). Such influences on N400 amplitudes have been obtained not only for written, spoken, and signed words (Kutas, Neville, & Holcomb, 1987), but also for other meaning-evoking stimuli, such as faces (Barrett & Rugg, 1989), objects (Barrett & Rugg, 1990), sounds (Van Petten & Rheinfelder, 1995) and mathematical symbols (Niedeggen, Rösler, & Jost, 1999).

Furthermore, N400 amplitudes decrease with repetition priming (Nagy & Rugg, 1989), and this repetition-induced decrease is modulated by lexical and semantic variables, namely being stronger for low frequency words (Rugg, 1990), and for words with many semantic features (Rabovsky, Sommer, & Abdel Rahman, 2012b). In addition, N400 amplitudes have been shown to vary as a function of lexical or semantic properties of single words. They are influenced by, for example, word frequency, with larger amplitudes for low frequency words (Van Petten & Kutas, 1990), and orthographic neighborhood size, with larger amplitudes for words with many orthographic neighbors (Holcomb, Grainger, & O'Rourke, 2002). Finally, N400 amplitudes are influenced by semantic richness (indicated by the number of semantic features or associates, word concreteness, or the depth of associated knowledge), with larger amplitudes for words with richer semantic representations (Holcomb, Kounios, Anderson, & West, 1999; Kounios & Holcomb, 1994; Laszlo & Federmeier, 2011; Müller, Dunabeitia, & Carreiras, 2010; Rabovsky, Sommer, & Abdel Rahman, 2012a, 2012c).

1.2. N400 theories and models

Despite the large body of data on N400 modulations, as yet there is no agreement on the specific mechanisms underlying this component. Various theories concerning the functional basis of the N400 have been proposed, suggesting that N400 amplitudes reflect the effort or difficulty involved in semantic memory access (Kutas & Federmeier, 2000; Van Berkum, 2009), the match or mismatch between expected and encountered semantic features (Paczynski & Kuperberg, 2012), lexical access (Lau, Phillips, & Poeppel, 2008), semantic integration/unification (Baggio & Hagoort, 2011; Brown & Hagoort, 1993), semantic binding (Federmeier & Laszlo, 2009), or semantic inhibition (Debruille, 2007). Thus, there are a variety of verbally-described theories of the N400's functional basis, most of them relating N400 amplitudes to an aspect of semantic processing.

The debate regarding the functional basis of the N400 could benefit from simulating the meaning-related N400 component in implemented computational models of meaning. Implemented computational models explicitly specify assumed mechanisms, and allow comprehensive exploration of the implications of theoretical assumptions (McClelland, 2009). Thus, relating N400 amplitude modulations to variations in measures of implemented computational models might advance a mechanistic understanding of the functional basis of the N400.

There are at least two possible approaches for gaining a better understanding of the N400 component by means of computational modeling. First, one might attempt to implement a neurally realistic model of the brain processes underlying the generation of this ERP component. In this spirit, Laszlo and Plaut (2012) presented a model of ERPs during word, nonword, and acronym reading. Their model incorporated neurophysiologically motivated constraints to enhance its neural plausibility and produce curves of total semantic activation that resembled the N400 component. Specifically, they separated excitation and inhibition (each unit could either be inhibitory or excitatory), reduced the number of inhibitory units, allowed inhibitory connections within but not between layers, and used a particular ("elbowed") inhibitory response function. Using arbitrary semantic representations, they compared mean activation of the semantic units for words, acronyms, pseudowords, and consonant strings, and found patterns of semantic activation that mirrored influences of orthographic neighborhood size on N400 amplitudes. Notably, in addition to simulating the N400, these authors also contributed to the literature by making more general points concerning how to achieve neurobiological plausibility when simulating ERPs with connectionist models.

1.3. Our modeling approach

There remain major challenges entailed by the complexities of the neural processes underlying the generation of scalp-recorded ERPs. Therefore, a second possible approach is to provide further understanding of the cognitive mechanisms underlying the N400 by directly relating variations in N400 amplitudes to variations in parameters produced by models of cognitive processes while leaving out the level of description of their complex neural realization and refraining from attempting to reproduce N400 morphology. This is the approach that we take, using a model that has successfully simulated a number of behavioral results in the semantic memory literature (Cree, McNorgan, & McRae, 2006; Mirman & Magnuson, 2008; O'Connor, Cree, & McRae, 2009). Following the common approach of describing the functional significance of ERP components depending on which factors modulate their amplitude, we investigated the functional significance of the N400 by examining whether variations in N400 amplitudes over a series of N400 paradigms would covary with a specific cognitive process (as simulated by an implemented model of meaning). This approach may entail losing some possibly interesting information with respect to neural realization. However, it allowed us to focus on our main goal, which is to understand the cognitive mechanisms underlying the N400 component, and thus to contribute to the theoretical debate about its functional basis.

We focused on connectionist models with error-driven learning because these have been successfully used to explain a range of phenomena in semantic cognition (Cree, McRae, & McNorgan, 1999; Cree et al., 2006; O'Connor et al., 2009; Rogers & McClelland, 2008). From a connectionist perspective, as discussed by Rogers and McClelland (2008), semantic cognition is considered to arise from activation flowing among simple processing units according to the strength of the weights connecting the units. Semantic knowledge is assumed to be stored in these connection weights which are gradually adjusted by experience. In response to a perceptual input such as a word or an object, it is assumed that the semantic system makes information available that is not directly present in the input. Thus, upon seeing a dog or a bird, a person with a fully developed semantic system can roughly anticipate how it would sound if the respective animal indeed made a sound, or what might happen if a cat approached it. Similarly, when reading sentences such as "I take my coffee with cream and..." contextual information interacts with semantic memory, generating anticipations of congruent semantic features which may facilitate the processing of possible sentence continuations. Semantic representations develop by adjusting the connection weights supporting these anticipations based on the difference between the anticipations and actual outcomes, resulting in more accurate anticipations as learning progresses. In this way, such implicit predictions gradually improve through experience.

In connectionist models, such learning corresponds to the activation of a specific input pattern (representing perceptual input), which results in activation flowing according to connection weights, producing a specific output pattern (representing implicit predictions) which is compared to a target pattern (representing actual outcomes). Connection weights are then adjusted based on network error, which is the difference between the model's generated output and the target (representing implicit predictions and actual outcomes, respectively).

For the present simulations, we used Cree et al.'s (2006) connectionist attractor model of word meaning (depicted in Fig. 1). The model has a two-layer architecture. The input layer consists of 30 abstract word form units, in which each word form is represented by a unique set of three of the 30 units, and is meant to represent either orthography (spelling) or phonology (how words sound). The output layer consists of 2526 semantic feature units that represent word meaning, based on representations for 541 object concepts that were empirically derived from McRae et al.'s (2005) semantic feature production norms. McRae et al. had participants list semantic features for 541 living and nonliving thing basic-level concepts, and retained all features that were produced by at least 5 of 30 participants. Across the 541 concepts, a total of 2526 distinct semantic features were produced. Thus, in the model, each word's meaning is represented as a distributed pattern across the 2526 semantic feature units (an activation value of 1 if a feature was listed for the respective concept in the norms by at least 5 participants, and 0 if not). Thus, the semantic representation for *dog*, for example, included features such as <has 4 legs>, <barks>, <has fur>, <has a tail>, <is domestic>, and so on. The word form input units are unidirectionally fully connected to the semantic units, and the semantic units are fully bidirectionally interconnected, with no self-connections. The word form units are not interconnected.

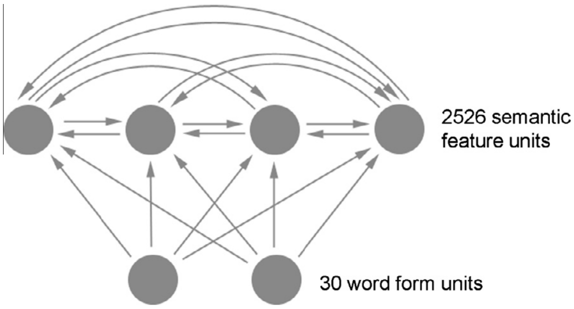

Fig. 1. Network architecture.

When a word form is presented to the model at the input layer, the model's task is to compute word meaning, that is, to activate the semantic features associated with that word form. It does so by gradually activating the corresponding semantic feature units via both the connections from the input units to the semantic units, and the direct interconnections between semantic feature units. Because it is an attractor network, this process happens over time. In the case of the present model, semantic representations settle over 20 so-called 'time ticks' or activation cycles (representing model time).

The model was trained using the recurrent backpropagation through time training algorithm (Pearlmutter, 1995). Thus, during each training trial, the input was a word form such as *dog*, and the model activated semantic features depending on the current connection strengths (which are random in the beginning, so that the activated features are initially random as well). Subsequently, the target (the correct semantic activation pattern) was presented to the model. For example, the target representation for *dog* includes activations of 1 for *dog* features such as <has 4 legs> and <barks>, with all features not included in the *dog* concept having a target activation of 0. The model then learns by adjusting the strength of all connections based on network error, obtained by calculating for each semantic unit the discrepancy between the activation produced by the model (varying between 0 and 1) and the target semantic activation (either 0 or 1). One training run through all 541 words from McRae et al.'s (2005) feature production norms is called a training epoch. After 20 training epochs, the model performed well in terms of activating the correct semantic feature units and deactivating the other feature units for all 541 words.

Including only abstract word form input units and semantic feature units, the model provides a simplified description of word recognition processes as a whole. However, the present research focuses on understanding the mechanisms underlying the N400 ERP component, and this focus is consistent with the fact that our model was designed primarily to provide insight into the semantic factors underlying word processing. That is, because the N400 seems to be a functionally specific electrophysiological indicator of semantic processing (Kutas & Federmeier, 2011), independent of input modality (e.g., visual or auditory) and domain[1] (e.g., words or pictures), it seems appropriate to use a model that emphasizes semantic computations.

Two measures taken from the model's semantic feature layer seem to be of particular interest in relation to the N400 component. The first is the total amount of semantic activation. This was the measure used by Laszlo and Plaut (2012). However, it is important to note that their network differed in several respects from ours (as described above), so that the activation results from the two models are not directly comparable. Total semantic activation corresponds to the sum of the activation of all semantic feature units (each ranging from 0 to 1), calculated at each of the 20 time ticks. Conceptually, this corresponds to the total amount of activation in the semantic system in response to a specific word. With respect to theories of the N400, total semantic activation may be related to the ease or difficulty of semantic access (Kutas & Federmeier, 2000), insofar as eliciting more semantic activity could be viewed as being more difficult. In addition, if predicted semantic features are activated in advance of the stimuli, then transient semantic activation can correspond to the match or mismatch between expected and encountered semantic features (Paczynski & Kuperberg, 2012) and the difficulty of semantic integration/unification (Baggio & Hagoort, 2011).

Second, semantic network error also is a promising candidate (we used cross-entropy error, which is described in Section 2.4). Network error is based on the difference, for each semantic unit, between the activation produced by the model and the correct semantic activation (1 for features that belong to the concept, and 0 for all other features). As noted above, network error is often conceptualized as implicit prediction error, with model-generated and correct output representing implicitly predicted and actual information respectively (Elman, 1990; McClelland, 1994; O'Reilly, Munakata, Frank, Hazy, & Contributors, 2012; Rogers & McClelland, 2008). Interestingly, McClelland (1994) discusses a relation between implicit prediction error in connectionist models and the N400 component. The idea that the neuro-cognitive system is "constantly predicting upcoming input and monitoring the consistency of the anticipated and actual outcomes" (Kutas, DeLong, & Smith, 2011, p. 202) has received a wave of support over the years (Bar, 2009; Barsalou, 2009; Clark, 2013; Friston, 2009; Furl, van Rijsbergen, Treves, Friston, & Dolan, 2007; Rao & Ballard, 1999; Schultz, Dayan, & Montague, 1997; Summerfield et al., 2006), and has often been related to event-related brain potentials (Chase, Swainson, Durham, Benham, & Cools, 2011; Cohen & Ranganath, 2007; Friston, 2005; Garrido, Kilner, Stephan, & Friston, 2009; Holroyd & Coles, 2002; Wacongne, Changeux, & Dehaene, 2012; Walsh & Anderson, 2012).

It is important to note that this type of implicit prediction and prediction error is not meant to correspond to any active, conscious, explicit prediction of specific items. The idea that is central to the current account is rather that the brain constantly extracts statistical regularities from the environment and builds an internal model of probability distributions in the environment to optimize processing. Constantly aiming to update this internal model, the brain is highly sensitive to mismatches between the internal model and the external world, which can be considered as implicit prediction errors. For instance, 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. Thus, in this sense of implicit prediction error, there does not need to be interdependences among subsequent time points for prediction to occur. Implicit prediction error is influenced not only by prior activation induced by the relation between subsequent time points, as for example in semantic priming paradigms, or in a sentence context, but also by connection strengths. Connection strengths are shaped by experience and hence by statistical regularities in the environment so that they reflect probability distributions, as in a higher probability for a high frequency than for a low frequency word. An alternative term for such implicit prediction error would be mismatch with Bayesian priors,[2] or Bayesian surprise (Clark, 2013; Doya, Ishii, Pouget, & Rao, 2007; Ostwald et al., 2012). Thus, in that sense, N400 amplitudes could be viewed as reflecting Bayesian surprise in the semantic system.

Indeed, N400 amplitude modulations in many paradigms appear to have one thing in common. Specifically, they are a function of the fit between the information that is implicitly anticipated based on statistical regularities across levels of representation as represented in semantic memory (semantic context, relations between words, frequency of occurrence of single words) and the actually presented information. Thus, we hypothesized that N400 amplitudes might reflect implicit prediction error in semantic memory in this sense of the term, represented by error values in a network model of meaning. This model-based account of the N400 in terms of implicit prediction error in the semantic system conceptually overlaps to various degrees with some of the verbally described theories of the N400. Specifically, the current account maps well onto the views that the N400 reflects the ease or difficulty of accessing features in semantic memory (Kutas & Federmeier, 2011), or the match or mismatch between

[1] There are, however, slight differences in the spatial distributions across the scalp (see e.g. Van Petten & Luka, 2006)

[2] The term "prior" may seem inappropriate, because relevant predictions depend on the current context and therefore may be viewed as "posterior" probabilities. However, in sequential Bayesian updating, even though context-dependent information may already be integrated in a prediction, this "posterior" relative to lacking contextual information becomes the new "prior" relative to the upcoming word. We believe that the N400 reflects the prediction error contained in the prior relative to the upcoming word.

expected and encountered semantic features (Paczynski & Kuperberg, 2012). We return to this issue in Section 4.

Because data on lexical decision performance is available for the N400 effects that we simulated, we also related both network error and semantic activation to behavioral performance. However, it seems important to note that unlike ERPs which reflect specific sub-processes occurring between stimulus presentation and response (such as semantic processes in the case of the N400), behavioral performance in lexical decision tasks presumably relies on a variety of sub-processes and mechanisms, some of which are not implemented in the present model. For instance, the model is lacking in the sense that the word form layer is abstract, and no attempt was made to simulate, for example, letter features, letters, phonemes, or statistical regularities within orthography or phonology. Furthermore, the model does not include components that simulate decision processes. Following Seidenberg and McClelland (1989), we assume that word reading involves orthographic, phonological, and semantic processes, and that lexical decision tasks require participants to establish appropriate decision criteria to discriminate between words and nonwords. These decision criteria can in principle involve any of the several dimensions along which words and nonwords differ (e.g., meaning, orthographic familiarity, phonological familiarity) and are presumably optimized according to the stimulus composition in an experiment. For instance, when the nonwords are highly orthographically atypical (e.g. *kzlpxr*), participants can base their decisions entirely on orthographic information (the presence or absence of familiar letter combinations), so that semantic variables may have little influence on the decisions. As an example, the richness of semantic representations typically facilitates lexical decisions (Pexman, Lupker, & Hino, 2002). However, using orthographically rather atypical nonwords, Rabovsky et al. (2012c) observed semantic richness effects on N400 amplitudes, indicating an influence of semantic richness on semantic processes, but these influences on semantic processes did not have any effects on lexical decision performance, presumably because participants adopted response criteria based on orthographic familiarity (for other examples of divergences between N400 amplitudes and behavioral responses please see e.g. Blackford, Holcomb, Grainger, & Kuperberg, 2012; Holcomb, Grainger, & O'Rourke, 2002).

However, whenever meaning-related decision criteria are useful to optimize performance in a given experimental situation (Seidenberg & McClelland, 1989), we assume that semantic computations as implemented in the present model feed into lexical decision processes. For example, because words have meanings but nonwords do not, high semantic activation can facilitate crossing a decision threshold to produce a "word" decision. Thus, we also examined the relation between measures taken from the model's semantic feature layer and behavioral performance in lexical decision tasks.

In summary, the primary goal of the present study was to test the hypothesis that N400 amplitudes reflect implicit prediction error in the semantic system, represented by network error values in an implemented model of meaning. To this end, we used Cree et al.'s (2006) model of word meaning to simulate seven N400 effects obtained in human empirical research on word processing. We examined the correspondence between the N400 and both semantic network error and semantic activation. Specifically, we simulated influences of semantic priming (Simulation 1), semantic richness (Simulation 2), word frequency (Simulation 3), repetition (Simulation 4), as well as influences of semantic richness on repetition effects (Simulation 5), influences of word frequency on repetition effects (Simulation 6), and influences of orthographic neighborhood size (Simulation 7).

2. The model

2.1. Network architecture

The network is depicted in Fig. 1. There are 30 word form input units that are intended to represent either orthography or phonology. These units are fully unidirectionally connected to 2526 directly interconnected semantic feature units representing word meaning. Directly interconnected feature units (without self-connections) were used so that the network naturally and transparently learns correlations among features, a variable that influences a number of semantic tasks (McRae, Cree, Westmacott, & de Sa, 1999; Tyler & Moss, 2001).

2.2. Training patterns

2.2.1. Word form units

Input patterns were abstract word form representations that can be interpreted as either spelling or sound. Each word's form was represented by activating a unique set of 3 of the 30 input units to 1, whereas the 27 other input units were turned off (0). Of the 91390 possible input patterns, 541 were assigned randomly (without replacement) to the 541 concept names from McRae et al.'s (2005) semantic feature production norms. Thus, other than precluding identical word form representations, there were no constraints on overlap among them. The representations were intended to capture the facts that word form representations overlap to varying degrees, and that the relation between word form and meaning is largely arbitrary for monomorphemic English words.

2.2.2. Semantic units

Semantic representations were based on McRae et al.'s (2005) semantic feature production norms, as described in Section 1. Note that we used semantic representations precisely as produced by human participants, so that there are no experimenter degrees of freedom in the feature-based representations. Finally, another advantage of using words/concepts from McRae et al.'s norms is that because a number of the experiments that are simulated in this article used stimuli from those norms, we were able to use the identical stimuli in our simulations.

2.3. Computation of word meaning

To compute a word's meaning, the corresponding word form was presented at the input layer and activation propagated to the semantic layer for 20 time ticks, segmented into 4 time steps, each consisting of 5 ticks. The activation pattern produced at the semantic layer was interpreted as the meaning of the word. Before presenting a word form, semantic units were set to randomly selected activation values between 0 and .1 so that the network began in a random initial state.

Input to a semantic unit was calculated as the activation of a sending unit multiplied by the weight of the connection from that sending unit. The net input x_j (at tick t) to unit$_j$ was calculated according to Eq. (1),

$$x_j^{[t]} = \tau\left(\sum_i s_i^{[t-\tau]} w_{ji} + b_j\right) + (1-\tau)x_j^{[t-\tau]} \tag{1}$$

where s_i is the activation of unit$_i$, and w_{ji} is the weight on the connection between unit$_i$ and unit$_j$. τ(tau) is a constant between 0 and 1 used to denote the duration of each time tick (0.2 in the present model). Thus, $x_j^{[t-\tau]}$ is the net input to unit$_j$ at the previous time tick. Each time tick is a subdivision of a time step, and consists of passing activation forward one step. Time ticks discretize and simulate continuous processing between time steps (Plaut, McClelland, Seidenberg, & Patterson, 1996). The activation of unit$_j$ at time t ($a_j^{[t]}$) was calculated using the sigmoidal activation function presented in Eq. (2), where $x_j^{[t]}$ is the net input to unit$_j$ from Eq. (1) (at time t).

$$a_j^{[t]} = \frac{1}{1 + e^{\left(-x_j^{[t]}\right)}} \tag{2}$$

Note that although we include all time ticks (except for the initial state at tick 0) in the simulation analyses presented below, we consider the first 3 ticks as being somewhat less informative than later ticks because they are primarily driven by the model moving away from the initial activation values. Also note that the initial summed activation at tick zero (which is approximately 126 for all but the semantic priming simulation) is not depicted in any figure, which instead begin with tick 1 when the summed semantic activation has begun to decrease. Including tick 0 would dramatically increase the scale of the figures' y-axes, and thus would distort the depiction of the experimental effects during the meaningful time ticks.

2.4. Training

The model was trained with the continuous backpropagation through time algorithm (Pearlmutter, 1995), using the rbp program of the PDPTool simulation package implemented in MATLAB (available at http://www.stanford.edu/group/pdplab/resources.html). At the beginning of training, weights were initialized to random values between ±.05. During each training epoch, all 541 concepts were presented in random order (exceptions in the training procedure for the frequency-sensitive model are reported in the frequency simulation section). On each training trial, the word form pattern was hard-clamped (i.e., activated at every time tick) at the input layer. For the first 10 time ticks, activation freely propagated to the semantic layer and within the semantic layer according to the current connection strengths in the model, and for the last 10 time ticks, the target semantic activation pattern was provided. This allows semantic patterns to settle gradually because the training regime does not constrain the network to compute the correct representation until tick 11. The difference between the model-generated semantic activation pattern and the target semantic activation pattern (network error) was used as a basis for adapting the connection weights (the connection weights between semantic units, and the weights between orthographic and semantic units, were modified by training). The learning procedure adjusts the strength of all connections in proportion to the extent to which this change reduces network error. Specifically, we used cross-entropy as the measure of network error. Cross-entropy error is more suitable than are the more frequently used squared-error measures for two reasons. First, during training, features can be considered as being either present (on) or absent (off), and intermediate states are understood as the probability that each feature is part of the concept. Thus, the activation of the semantic layer represents a probability distribution that is used in computing cross entropy (Plunkett & Elman, 1997). Second, cross-entropy error is beneficial for sparse networks; the present network is sparse because only 5–21 of the 2526 feature units should be on for each concept. Cross entropy produces large error values when a unit's activation is on the incorrect side of .5. Because one possible way for the present sparse network to reduce error is to turn off all semantic feature units, punishing it for incorrectly turning units off allows the model to more easily change these units' states. Cross-entropy error (E), averaged over the last two time steps (10 ticks) was computed as in Eq. (3),

$$E = \tau\frac{\sum_{t=10}^{19}\sum_{j=0}^{2526} d_j \ln(a_j) + (1-d_j)\ln(1-a_j)}{10} \tag{3}$$

where d_j is the desired target activation for unit$_{j,}$ and a_j is the unit's observed activation.

The learning rate was .01 and momentum of .9 was added after the first 10 training epochs. Because we assume that humans do not compute absolutely perfect semantic representations (there is a bit of noise), training was stopped after 20 epochs. At that point, over 99% of all features for all 541 concepts were correct in the sense of being activated above .7 for units intended to be on, and below .3 for units intended to be off.

During the simulations involving repetition effects (Simulations 4, 5, and 6), learning was left operative, using the same learning rate of .01 as in initial training. It is important to note that having learning 'on' vs. 'off' does not introduce any difference in network performance measures; the difference between 'on' and 'off' does not influence current error or activation values; influences are apparent only in subsequent simulation trials. To maintain consistency with the simulations that did not include repetition as a factor, rather than using on-line learning (i.e., weight updating after each item), we trained all items from the first presentation together as a group (batch training) between the first and second presentations of items. The network did not learn about the sequential order of the items because, as noted above, semantic units were reset to randomly selected activation values between 0 and .1 at the beginning of each trial.

It is also important to note that having learning ‘on’ is not assumed to reflect active instructed learning. Instead, it corresponds to implicit adaptation processes assumed to take place automatically as a result of stimulus processing (also see Stark & McClelland, 2000). Leaving learning ‘on’ may influence subsequent implicit prediction error via the adaptation of connection weights and thus, conceptually, via adaptation of representations of occurrence probability.

3. Simulations

We present seven simulations covering a number of variables known to influence N400 amplitudes. We investigated the correspondence between influences of these variables on N400 amplitudes, as reported in the literature, and influences of the same variables on two model measures, semantic network error, and total semantic activation. The most plausible initial assumption seems to be that larger values of a specific measure (error or activation) correspond to larger N400 amplitudes. The results are presented for network error on the left, and total semantic activation on the right. A summary of the simulation results is presented in Table 1. To foreshadow those results, in all seven simulations, higher network error corresponded to larger N400 amplitudes. On the other hand, higher semantic activation corresponded to larger N400 amplitudes in one simulation only. Thus, the reverse relationship, lower activation corresponding to larger N400 amplitudes, was observed in six of the seven simulations.

Because many variables known to influence N400 amplitudes also influence behavioral performance in lexical decision tasks, data on lexical decision performance are available for the simulated studies. Therefore, we also examined the correspondence between lexical decision performance and both semantic network error and activation. Although it is not clear how semantic network error could be used as a basis to distinguish words from nonwords because nonwords do not have semantic representations (meaning), error for words has often been related to the process of settling into a representation and hence decision latencies for different types of words, with lower error corresponding to response facilitation. In addition, higher semantic activation may correspond to facilitated responses, because in lexical decision tasks, words have meaning whereas nonwords do not, so that increased semantic activation may help crossing a threshold to produce a ‘word’ decision. To foreshadow the results, lower error corresponded to facilitated responses in five of six simulations, whereas higher activation corresponded to facilitated responses in six of six simulations (for the seventh simulation, concerning orthographic neighborhood effects, the empirical evidence is controversial; see below for further discussion).

3.1. Simulation 1: Semantic priming

Semantic priming paradigms frequently have been used to study the organization of semantic memory (see Hutchison, 2003; Lucas, 2000, for reviews). Although there are variations in procedures, in a typical priming study, a prime (e.g., *truck*) is presented for a short period of time, followed by a target (*van*). Participants read the prime, and then respond to the target, often by making a lexical decision. The related condition is compared to an unrelated condition in which the target is preceded by an unrelated prime (*shirt–van*). A number of studies have found reduced N400 amplitudes to target words following semantically related as compared to unrelated primes (Bentin et al., 1985; Kutas & Federmeier, 2011), as well as facilitated responses for related targets (McRae & Boisvert, 1998).

3.1.1. Methods

To simulate semantic priming, the model was presented with 36 related and unrelated word pairs used by McRae and Boisvert (1998), with the unrelated pairs created by

Table 1
Summary of simulations. The second line for each simulation indicates the human data (N400 amplitudes or behavioral measures, i.e., response times and error rates) with which the model patterns. In parentheses are the time ticks during which the effect is significant in the simulation.

Simulation	Human N400	Human behavior	Model error	Model activation
Semantic Priming	Related < unrelated	Related < unrelated	Related < unrelated (1–10) N400 & behavior	Related > unrelated (3–6) Behavior
Semantic Richness	High > low	High < low	High > low (1–7) N400	High > low (6–20) N400 & behavior
Frequency	High < low	High < low	High < low (4–20) N400 & behavior	High > low (6–7) Behavior
Repetition	Repeated < initial	Repeated < initial	Repeated < initial (1, 4–20) N400 & behavior	Repeated > initial (4–10) Behavior
Semantic Richness × Repetition	Interaction	Interaction	Interaction[a] (2, 4–20) N400 & behavior	Interaction[b] (5–10) Behavior
Frequency × Repetition	Interaction	Interaction	Interaction[c] (6–20) N400 & behavior	Interaction[d] (8–10, 15–17, 19–20) Behavior
Orthographic Neighbors	Many > few	?	Many > few (8–20) N400	Many < few (1, 2, 4–8) –

[a] Stronger repetition-induced reduction for words with many features.
[b] Stronger repetition-induced increase for words with many features.
[c] Stronger repetition-induced reduction for low frequency words.
[d] Stronger repetition-induced increase for low frequency words.

re-arranging the primes and targets from the related pairs. According to McRae et al.'s (2005) norms, the target and related prime shared 6.6 features on average (range: 2–14), whereas the target and unrelated prime shared only 0.9 features on average (with shared features typically being general features such as <is large>). To simulate each trial, activation of semantic units were initialized to random values between 0 and .1, the prime's word form was presented for 20 ticks, directly followed by presentation of the target word, for 20 ticks as well. Thus, when the target's word form was presented initially to the network, activation in the semantic layer corresponded to the settled activation pattern for the prime.

3.1.2. Results

The results of Simulation 1 are presented in Fig. 2. Semantic activation in Fig. 2 is the total semantic activation during the processing of the prime, and then the target. Because related primes were shuffled to create the unrelated prime-target pairs, semantic activation is nearly identical for the related and unrelated conditions during prime processing (with small deviations due to differences in random starting configurations). In contrast, the network error that is presented in Fig. 2 during prime processing is relative to the target word. That is, when *van* was the target, *truck* was the related prime, and *shirt* was the unrelated prime, error is presented relative to *van*. This provides a measure of the distance between the current activation in the network and the target word's semantic representation throughout a prime-target trial. This distance is shorter (network error is lower) for related compared to unrelated primes during prime processing because targets share a greater number of features with their related primes.

Network error and total semantic activation during target word processing were submitted to separate repeated measures analyses of variance. The two within-items independent variables were relatedness (semantically related vs. unrelated prime) and time (ticks 21–40). Thus, the analyses began one tick after the target word form was input to the network.

Analyses of network error to target words revealed lower error for semantically related targets, $F(1,35) = 72.16$, $p < .001$, $\eta^2 = .67$. The relatedness effect changed over time, as shown by a relatedness by time interaction, $F(19,665) = 88.36$, $p < .001$. Planned comparisons revealed significantly lower network error for related targets from ticks 21 to 30 (all Fs > 8.82, all $ps < 01$).

Analyses of activation showed greater total semantic activation for semantically related targets, $F(1,35) = 6.03$, $p < .05$, $\eta^2 = .15$. Relatedness again interacted with time, $F(19,665) = 3.21$, $p < .05$. Activation was marginally higher for related targets at ticks 22 and 27 (Fs > 3.95, $ps < .10$), and significantly higher from tick 23 to 26 (Fs > 5.52, $ps < .05$).

3.1.3. Discussion

Consistent with N400 amplitudes (Bentin et al., 1985; Kutas & Federmeier, 2011), network error was reduced for targets presented after semantically related as compared to unrelated primes. On the other hand, relatively early during the time course of computing word meaning, semantic activation was larger for related as compared to unrelated targets, which is opposite to well-established N400 results. Instead, the results may be taken to suggest that higher semantic activation (as observed in our model) could contribute to facilitating behavioral responses to semantically related as compared to unrelated targets (McRae & Boisvert, 1998).

3.2. Simulation 2: Semantic richness

N400 amplitudes are larger for words with richer semantic representations, such as concrete words (Holcomb et al., 1999; Kounios & Holcomb, 1994; Kounios et al., 2009; West & Holcomb, 2000), words with a higher number of semantic features (Amsel, 2011; Rabovsky et al., 2012c; but see Kounios et al., 2009), words with a higher number of associates (Laszlo & Federmeier, 2011; Müller et al., 2010), and newly learned words associated with in-depth as compared to minimal semantic information (Rabovsky et al., 2012a). A number of studies have shown facilitated behavioral responses for words

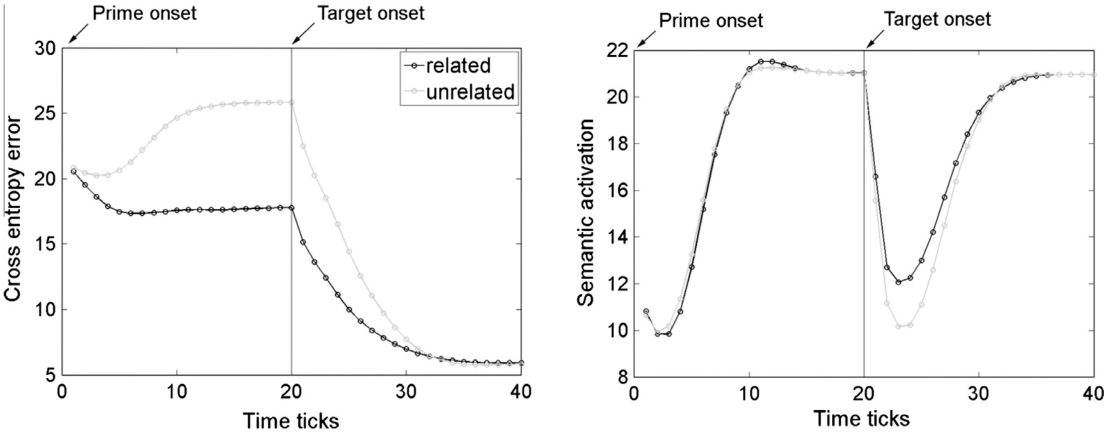

Fig. 2. Influences of semantic priming on network error (left) and activation (right).

with a greater number of semantic features (Grondin, Lupker, & McRae, 2009; Pexman et al., 2002), words with more associates (Dunabeitia, Aviles, & Carreiras, 2008) and concrete as compared to abstract words (e.g. Kounios & Holcomb, 1994).

3.2.1. Methods

We simulated semantic richness effects by presenting the model with Rabovsky et al.'s (2012c) stimuli comprised of 80 words with a high number of features ($M = 16.1$) and 80 words with a low number of features ($M = 9.3$), according to McRae et al.'s (2005) norms. Stimulus groups did not differ in terms of the number of associates, familiarity, concreteness, length, word frequency, or bigram frequency, phonological neighbors, phonemes and syllables as well as the number of orthographic neighbors and overlap in the model's word form representations (i.e. the number of word forms differing in a single input unit).

3.2.2. Results

The results of Simulation 2 are presented in Fig. 3. Error and activation were submitted to separate mixed analyses of variance with richness (high vs. low) as a between-items variable and time (ticks 1–20) as a within-items variable.

Error was higher for words with many as compared to few semantic features, $F(1,158) = 130.09$, $p < .001$, $\eta^2 = .31$. This richness effect interacted with time, $F(19,3002) = 179.78$, $p < .001$. Planned comparisons revealed that the richness effect was significant at ticks 1–7 (all Fs > 6.73, all ps < .05).

Semantic activation was higher for words with many as compared to few semantic features, $F(1,158) = 250.99$, $p < .001$, $\eta^2 = .19$. Richness interacted with time, $F(19,3002) = 95.83$, $p < .001$, with a significant richness effect at ticks 6–20 (all Fs > 8.52, $p < .01$).

3.2.3. Discussion

Both network error and activation were higher for words with many as compared to few semantic features. Thus, both measures could account for larger N400 amplitudes for words with richer semantic representations as reported in the literature (Amsel, 2011; Holcomb et al., 1999; Kounios & Holcomb, 1994; Kounios et al., 2009; Laszlo & Federmeier, 2011; Müller et al., 2010; Rabovsky et al., 2012a, 2012c; West & Holcomb, 2000). In addition, higher activation (but not higher error) could also account for the observed behavioral facilitation for words with richer semantic representations (Dunabeitia et al., 2008; Grondin et al., 2009; Pexman, Hargreaves, Siakaluk, Bodner, & Pope, 2008; Pexman, Holyk, & Monfils, 2003; Pexman et al., 2002).

3.3. Simulation 3: Word frequency

N400 amplitudes are larger for low frequency as compared to high frequency words (e.g. Barber, Vergara, & Carreiras, 2004; Rabovsky, Alvarez, Hohlfeld, & Sommer, 2008; Van Petten & Kutas, 1990). In addition, behavioral responses are facilitated for high frequency words (Forster & Chambers, 1973). From a connectionist perspective, word frequency effects reflect variability in the amount of training for the respective words.

3.3.1. Methods

To simulate influences of lexical frequency, the amount of training for every word was determined by its lexical frequency. Specifically, the number of occurrences of each of the 541 words per training epoch corresponded to the natural logarithm of its frequency in the British National Corpus (BNC). Please note that due to the different training procedure, the fine-grained details of the results, specifically the size of the effects, are not directly comparable with the other simulations. After training, we presented the network with 70 high frequency words, with mean BNC frequency of 3704 [mean ln(BNC) was 7.7], and 70 low frequency words, with mean BNC frequency of 66 [mean ln(BNC) was 3.8]. Stimuli were selected from McRae et al. (2005), and the high and low frequency words did not differ in terms of the number of semantic features, intercorrelational density, word length, as well as the number of orthographic neighbors and overlap in the model's word form representations.

3.3.2. Results

The results of Simulation 3 are presented in Fig. 4. Network error and activation were submitted to mixed analyses of variance with word frequency (high vs. low)

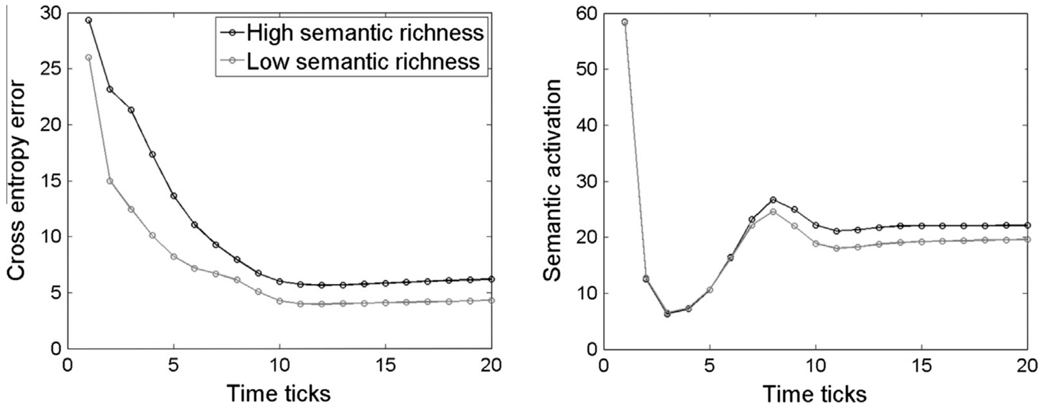

Fig. 3. Influences of semantic richness on network error (left) and activation (right).

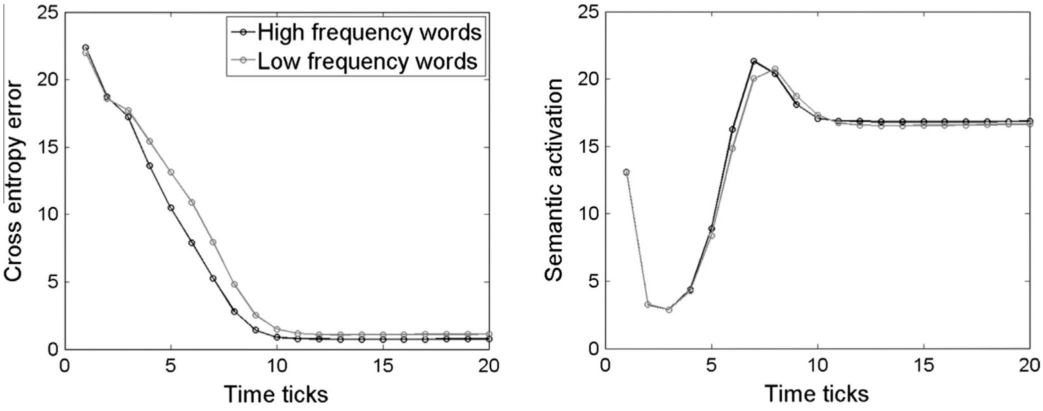

Fig. 4. Influences of word frequency on network error (left) and activation (right).

as a between-items variable and time (ticks 1–20) as a within-items variable.

Network error was significantly smaller for high frequency words, $F(1,138) = 10.19$, $p < .01$, $\eta^2 = .07$. Frequency interacted with time, $F(19,2622) = 7.91$, $p < .01$, with significant frequency effects extending from ticks 4 to 20 (all Fs > 5.01, all ps < 05).

For total semantic activation, there was no significant main effect of frequency ($F < 1$). However, frequency interacted with time, $F(19,2622) = 2.14$, $p < .05$. At ticks 6 and 7, there was higher activation for high than for low frequency words (Fs > 4.91, ps < .05, η^2s > .03). The frequency effect was marginally significant at tick 5, $F(1,138) = 3.06$, $p = .08$.

3.3.3. Discussion

In good correspondence with reduced N400 amplitudes for high frequency words (Barber et al., 2004; Rabovsky et al., 2008; Van Petten & Kutas, 1990), error was smaller for high frequency as compared to low frequency words. On the other hand, semantic activation was larger for high frequency words at ticks six and seven. Thus, there was little effect of frequency on semantic activation, and the differences that did occur were in the opposite direction of N400 amplitudes. If anything, the results are more in line with the possibility that higher semantic activation (as observed in our model) may contribute to behavioral facilitation for high frequency words (Forster & Chambers, 1973).

To obviate possible concerns that the word frequency effect might be driven by feature frequency rather than word frequency itself, we calculated the mean feature frequency of each word, that is, the number of words (concepts) in the training set in which a feature occurs times the frequency of those words, averaged across all features in a word. Feature frequency was actually higher for the low frequency words ($M = 155$) than for the high frequency words ($M = 124$), $t(69) = -3.32$, $p < .01$. Therefore, any possible confounding influences of feature frequency would presumably oppose the influence of word frequency. As there seems to be a relatively straightforward relationship between more training (i.e., higher frequency) and lower error, in the sense that training typically reduces error, we are confident that the obtained frequency effects on network error were caused by word frequency rather than feature frequency.

Finally, we note two factors that influence the size of the word frequency effects on network error. The first is the amount of training. Early on in training, word frequency effects are substantial, because the connections involved in the high frequency words are much better adapted than those involved in the low frequency words. However, later during training, low frequency words begin to catch up. This is because, on average, the high frequency words more quickly reach a point at which they induce relatively low error values so that error-driven learning produces smaller weight changes. In the meantime, low frequency words continue to produce relatively large error values, which induce larger changes in connection weights – they benefit more from each additional training trial, so that they catch up to higher frequency words to some degree. The second factor is the implementation of frequency. In the present simulation, we used the natural logarithm of frequency in the British National Corpus [ln(BNC)] as a proxy. If we had used the actual frequency in the BNC instead of the natural logarithm, the frequency effect would have been much larger (due to 3704 vs. 66 occurrences per epoch instead of 7.7 vs. 3.8). It seems difficult to know exactly which frequency difference in a model with only 541 words at which point in training corresponds to the difference between BNC frequency values of 3704 and 66 in proficient adult human readers who are exposed to a much wider and more dynamic range of stimuli. However, we were primarily interested in the direction of the effects, that is, which model measure was influenced in the same direction as N400 amplitudes, rather than their absolute magnitude. And the direction of the effects, that is, higher frequency producing lower error values, depends on basic learning principles in connectionist models.

3.4. Simulation 4: Repetition

N400 amplitudes are reduced by repetition (Kiefer, 2005; Kounios & Holcomb, 1994; Kutas & Federmeier, 2011; Nagy & Rugg, 1989; Rabovsky et al., 2012b; Rugg, 1990; Sim & Kiefer, 2005). In addition, repetition facilitates behavioral responses (Scarborough, Cortese, & Scarborough, 1977). In connectionist models, repetition effects have been simulated as the influence of one additional training trial on the respective word (Stark & McClelland, 2000).

3.4.1. Methods

We simulated repetition effects by presenting 160 concrete nouns (used by Rabovsky et al., 2012b) twice to the model, in two successive blocks. As noted above, in contrast to the other simulations reported herein, learning was left on throughout (learning rate = .01). We then compared error and semantic activation between the first presentation and the repetition.

3.4.2. Results

The results of Simulation 4 are presented in Fig. 5. Network error and activation were submitted to separate repeated-measures analyses of variance with repetition (first presentation vs. repetition) and time (ticks 1–20) as within-items variables.

Error was reduced for repetitions as compared to first presentations, $F(1,159) = 75.50$, $p < .001$, $\eta^2 = .32$. Furthermore, there was a significant interaction between repetition and time, $F(19,3021) = 68.32$, $p < .001$. Error was reduced for repetitions at tick 1 and at ticks 4–20 (all $Fs > 10.31$, $ps < .01$).

Total semantic activation was higher for repetitions as compared to first presentations of the stimuli, $F(1,159) = 5.07$, $p < .05$, $\eta^2 = .03$. The repetition effect also interacted with time, $F(19,3021) = 99.68$, $p < .001$. From ticks 4 to 10, activation was significantly higher for repetitions as compared to first presentations ($Fs > 25.57$, $ps < .001$). An effect in the opposite direction was obtained during the first two ticks ($Fs > 275.47$, $ps < .001$), although because this is the first two time ticks, it is unlikely to be meaningful. Due to the reversal in the direction of effects, there was no difference at tick 3 ($F < 1$).

3.4.3. Discussion

Network error corresponded with N400 amplitudes in that it was smaller for repetitions as compared to first presentations (Kiefer, 2005; Kounios & Holcomb, 1994; Nagy & Rugg, 1989; Rabovsky et al., 2012b; Rugg, 1990; Sim & Kiefer, 2005). In contrast, the results for activation are opposite to the N400 results. The extreme repetition effects on activation during the first two ticks is not highly informative because they presumably are driven primarily by the model moving away from the initial state activations of the feature units. During the rising edge of activation and around the activation maximum (ticks 4–10), activation was higher for repetitions as compared to first presentations, which is opposite to well-established N400 results, and instead seems more in line with the idea that higher activation (as observed in our model) can facilitate behavioral responses, resulting in facilitated performance for repeated words (e.g. Scarborough et al., 1977).

It should be noted that the size of the repetition effect depends partly on the amount of training. Because the present model was not trained to asymptote (see Methods), there remained an opportunity for a substantial influence of one additional presentation (i.e., the most recent weight changes). If the model was trained initially for a greater number of epochs, repetition effects would have been smaller. Furthermore, repetition effects potentially depend on the number of repeated words in the sense that the size of the repetition effects might differ depending on whether the changes in the connection weights for the recently experienced words (concepts) partly interfere or are consistent with one other. However, we were primarily interested in the direction of the effects rather than their absolute magnitude, and we are confident that the direction, that is, an additional training trial entailing reduced error values, is stable across such variations.

3.5. Simulation 5: Influences of semantic richness on repetition effects

A recent study has found semantic richness to enhance repetition effects, with larger repetition-induced reductions of N400 amplitudes and error rates for words with many as compared to words with few semantic features (Rabovsky et al., 2012b). Similar results have been

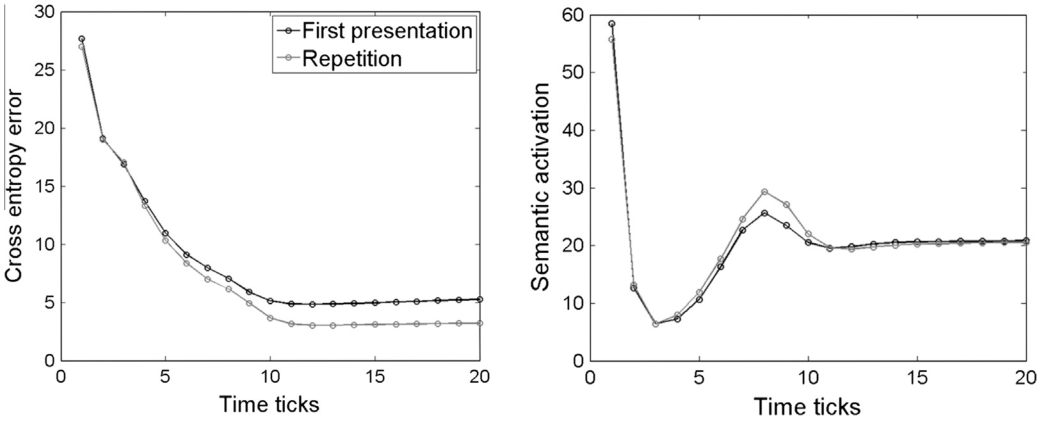

Fig. 5. Influences of repetition on network error (left) and activation (right).

obtained for concrete vs. abstract words (Kounios & Holcomb, 1994).

3.5.1. Methods

To simulate the influence of semantic richness on repetition effects, we used the 80 words with many and 80 words with few semantic features from Rabovsky et al. (2012b) that were also used in Simulation 2. The stimuli were presented twice to the model, in two successive blocks, with learning left operative. We then compared repetition effects on network error and activation for words with many vs. few semantic features.

3.5.2. Results

The results of Simulation 5 are presented in Fig. 6. Network error and activation were submitted to separate mixed analyses of variance with richness (high vs. low) as a between-items variable, and repetition (first presentation vs. repetition) and time (ticks 1–20) as within-items variables.

Overall, error was reduced for the repeated presentation of words, $F(1,158) = 80.89$, $p < .001$. Importantly, there was a significant interaction between richness and repetition, $F(1,158) = 12.36$, $p < .01$, $\eta^2 = .07$, indicating stronger repetition-induced reduction for words with many semantic features, $F(1,79) = 52.95$, $p < .001$; $\eta^2 = .40$, as compared to words with few semantic features, $F(1,79) = 28.73$, $p < .001$, $\eta^2 = .27$.

The interaction between richness and repetition varied across time, giving rise to a significant three-way interaction, $F(19,3002) = 4.49$, $p < .05$). Further analyses showed that the interaction between richness and repetition was significant at tick 2, as well as from ticks 4 to 20 (all *F*s > 6.21, all *p*s < .05).

Semantic activation was larger for repeated words as compared to their initial presentation, $F(1,158) = 5.56$, $p < .05$. There was a significant interaction between richness and repetition, $F(1,158) = 16.40$, $p < .001$, $\eta^2 = .09$. This interaction varied across time ticks, giving rise to a significant three-way interaction between richness, repetition and time, $F(19,3002) = 11.45$, $p < .001$.

As can be seen in Fig. 6, this three-way interaction was mainly due to two groups of time ticks differing concerning the nature of the interaction between richness and repetition. Relatively early on, during the rising edge of activation and around the maximum (ticks 5–10), repetition induced an increase in activation that was more pronounced for words with many features than for words

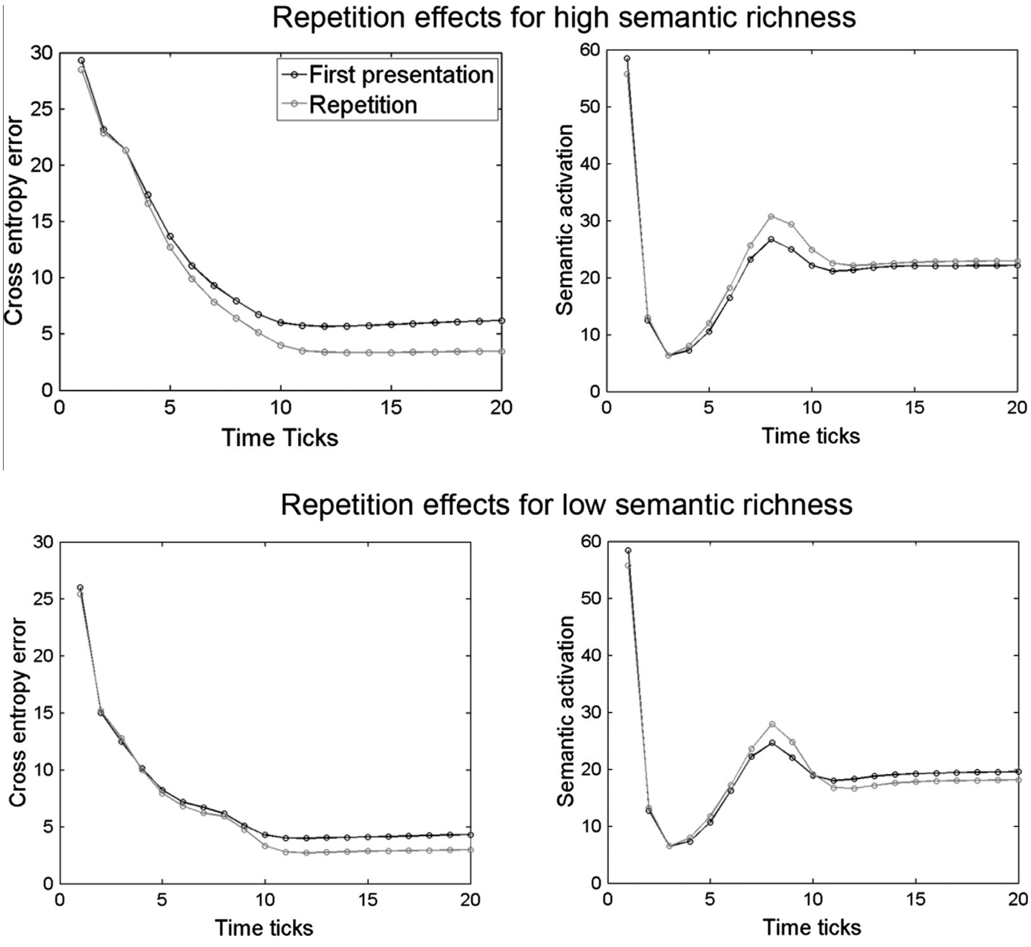

Fig. 6. Influences of repetition on network error (left) and activation (right), for words with either high semantic richness (top) or low semantic richness (bottom).

with few features. Analyses of ticks 5–10 revealed an influence of repetition, $F(1,158) = 142.30$, $p < .001$, and a richness by repetition interaction, $F(1,158) = 10.30$, $p < .01$. This interaction was due to a larger repetition effect for words with many features, $F(1,79) = 87.85$, $p < .001$, $\eta^2 = .53$, than for words with few features, $F(1,79) = 54.65$, $p < .001$, $\eta^2 = .41$.

On the other hand, later on, repetition slightly increased activation for words with many features, while decreasing activation for words with few features. Analyses of ticks 11–20 revealed no main effect of repetition, $F(1,158) = 1.53$, $p > .2$, but a significant interaction between richness and repetition, $F(1,158) = 18.00$, $p < .001$. For words with many features, the slight repetition-induced increase was marginally significant, $F(1,79) = 2.93$, $p < .1$. This was qualified by an interaction between repetition and time, $F(9,711) = 7.62$, $p < .01$, with further tests showing that repetition was significant at tick 11, $F(1,79) = 7.93$, $p < .01$, and a trend at ticks 17–20 (Fs > 2.80, ps < .10). For words with few features, there was actually a decrease in activation from the initial to repeated presentations, $F(1,79) = 32.66$, $p < .001$.

3.5.3. Discussion

Consistent with N400 amplitudes (Kounios & Holcomb, 1994; Rabovsky et al., 2012b), network error showed an enhanced repetition-induced reduction for words with many features as compared to words with few features. Activation values corresponded less well with N400 amplitudes. During the rising edge of the activation and around the maximum, activation showed an enhanced repetition-induced increase for words with many as compared to words with few semantic features. Although the enhancement of the repetition effect for words with many features is in line with findings for the N400, the overall direction of the repetition effect, namely a repetition-induced increase, is opposite to repetition effects on the N400. Instead, if higher semantic activation can facilitate performance, the results for the activation measure could account for the enhanced repetition-depended facilitation of lexical decisions for words with many features (Rabovsky et al., 2012b).

3.6. Simulation 6: Influences of frequency on repetition effects

Repetition-induced N400 amplitude reductions have been found to be enhanced for low frequency as compared to high frequency words (Rugg, 1990; Young & Rugg, 1992). In addition, in behavioral measures, low frequency words have also shown stronger repetition effects (Forster & Davis, 1984; Norris, 1984).

3.6.1. Methods

We used the model from Simulation 3 in which the amount of training depended on lexical frequency according to the British National Corpus. Stimuli were the same 70 high frequency and 70 low frequency words (mean BNC frequency of 3704 vs. 66, mean ln(BNC) of 7.7 vs. 3.8), used in Simulation 3. Stimuli were presented twice to the model, in two successive blocks, with learning left operative, and repetition effects were compared between high and low frequency words.

3.6.2. Results

The results of Simulation 6 are presented in Fig. 7. Network error and activation were submitted to separate mixed analyses of variance with frequency (high vs. low) as a between-items variable, and repetition (first presentation vs. repetition) and time (ticks 1–20) as within-items variables.

Overall, error was reduced for repeated words, $F(1,138) = 114.88$, $p < .001$. Importantly, this repetition effect was modulated by frequency, $F(1,138) = 13.83$, $p < .001$, $\eta^2 = .09$. The interaction resulted from stronger repetition-induced reduction for low frequency words, $F(1,69) = 74.97$, $p < .001$, $\eta^2 = .52$, as compared to high frequency words, $F(1,69) = 40.17$, $p < .001$, $\eta^2 = .37$. There also was a three-way interaction among repetition, frequency, and time, $F(19,2622) = 7.37$, $p < .001$. The frequency by repetition interaction was marginal at tick 5, $F(1,138) = 3.77$, $p = .054$, and significant from ticks 6 to 20 (all Fs > 7.67, all ps < .01).

Semantic activation was enhanced for repeated words, $F(1,138) = 62.80$, $p < .001$. Importantly, this repetition effect was modulated by frequency, $F(1,138) = 5.62$, $p < .05$, $\eta^2 = .04$. Further analyses showed a stronger repetition-induced increase for low frequency words, $F(1,69) = 36.70$, $p < .001$, $\eta^2 = .35$, than for high frequency words, $F(1,69) = 27.74$, $p < .001$, $\eta^2 = .29$. There was a three-way interaction among repetition, frequency, and time, $F(19,2622) = 3.59$, $p < .01$. Repetition interacted with frequency at ticks 8–10, as well as at ticks 15, 16, 17, 19, and 20 (all Fs > 4.15, all ps < .05), and it was marginally significant at ticks 12, 14, and 18 (all Fs > 2.88, all ps < .10).

3.6.3. Discussion

Consistent with N400 amplitudes (Rugg, 1990; Young & Rugg, 1992), network error decreased with repetition, and this repetition-induced decrease was stronger for low frequency as compared to high frequency words.

In contrast, semantic activation corresponded less well with N400 amplitudes. Although frequency interacted with repetition, with larger effects for low frequency words, the overall direction of this repetition effect, namely an increase of activation for repeated words, is opposite to N400 data (Rugg, 1990; Young & Rugg, 1992). On the other hand, semantic activation was consistent with the observation of stronger repetition-induced behavioral facilitation for low frequency words (Forster & Davis, 1984; Norris, 1984).

3.7. Simulation 7: Influences of orthographic neighborhood size

N400 amplitudes are larger for words with many orthographic neighbors (N; i.e., the number of words that can be obtained from the target word by exchanging a single letter, preserving letter position; Holcomb, Grainger, & O'Rourke, 2002; Laszlo & Federmeier, 2011). The present model focuses on semantic representations and is not optimally suited for simulating orthographic neighborhood

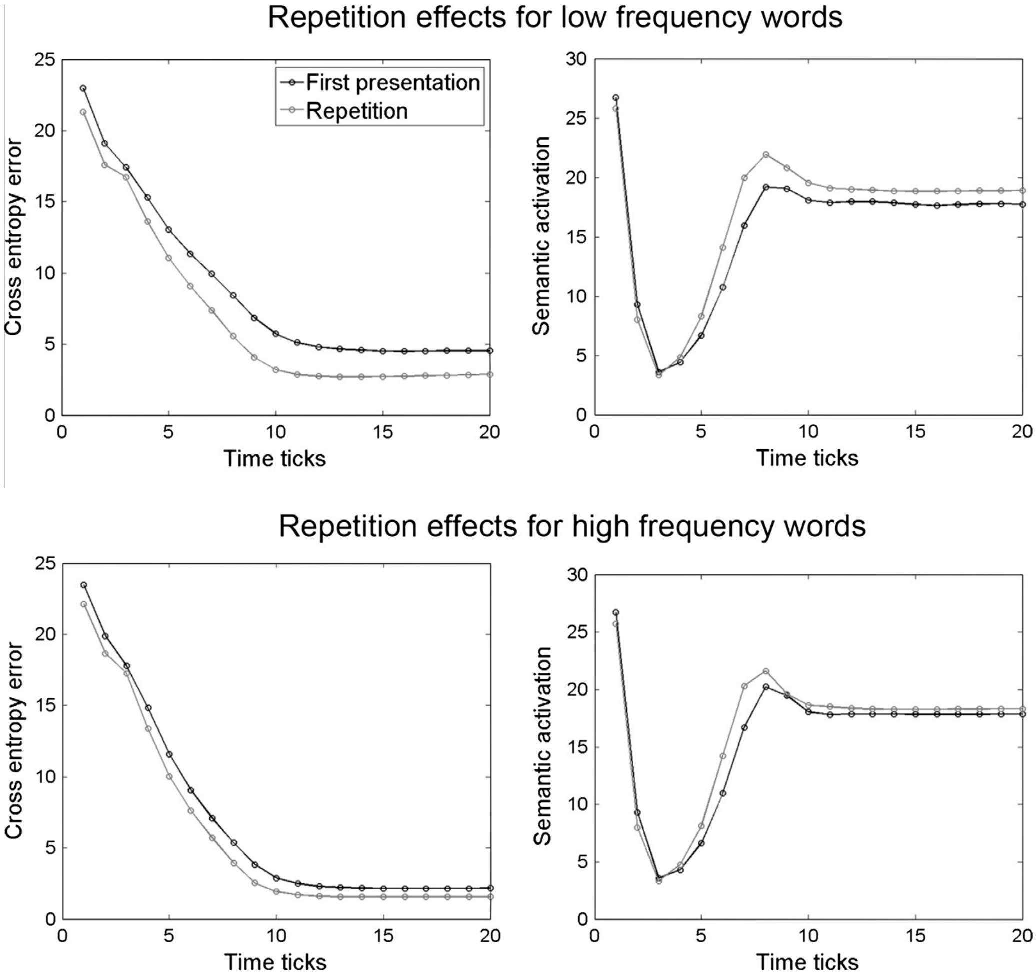

Fig. 7. Influences of repetition on network error (left) and activation (right), for words with either low lexical frequency (top) or high lexical frequency (bottom).

effects because statistical orthographic regularities are not encoded in the word form input layer. However, because the only previous model of the N400 specifically focused on these effects (Laszlo & Plaut, 2012), we also simulated influences of orthographic neighborhood size.

Whereas orthographic neighborhood size consistently has been found to increase N400 amplitudes (Holcomb, Grainger, & O'Rourke, 2002; Laszlo & Federmeier, 2011), its influence on behavioral responses is inconsistent across studies. Holcomb et al. (2002) found shorter decision latencies for words with a greater number of orthographic neighbors. However, Coltheart, Davelaar, Jonason, and Besner's (1977) original investigation of orthographic neighborhood effects found an inhibitory effect for nonword decisions, and no influence on word decisions. Other studies have obtained conflicting results, with Andrews (1989) reporting a facilitatory effect of orthographic neighborhood size, but only for low frequency words, whereas Grainger, Oregan, Jacobs, and Segui (1989) found inhibitory effects. Later attempts to resolve the conflict revealed that orthographic neighborhood effects on behavioral performance depend on many factors including the relative frequency of the orthographic neighbors and the type of overlap (e.g. Andrews, 1997; Ziegler & Perry, 1998). Attempting to reconcile this set of behavioral data is beyond the scope of the present study, which is primarily concerned with influences on N400 amplitudes.

3.7.1. Methods

We calculated the amount of overlap among word form representations. Specifically, as a model estimate of orthographic neighborhood size, for each word form representation, we calculated the number of other word forms with which it shared two input units and thus differed by only a single input unit. Word form representations differing in a single input unit were considered orthographic neighbors. The number of orthographic neighbors ranged from 3 to 21 ($M = 10.9$). For the simulation, we then presented the model with the 70 words that had the largest orthographic neighborhood ($\geqslant 15$ neighbors) and the 70 with the smallest ($\leqslant 7$ neighbors). These groups did not differ in terms of the number of semantic features.

3.7.2. Results

The results of Simulation 7 are presented in Fig. 8. Error and activation were submitted to separate mixed analyses

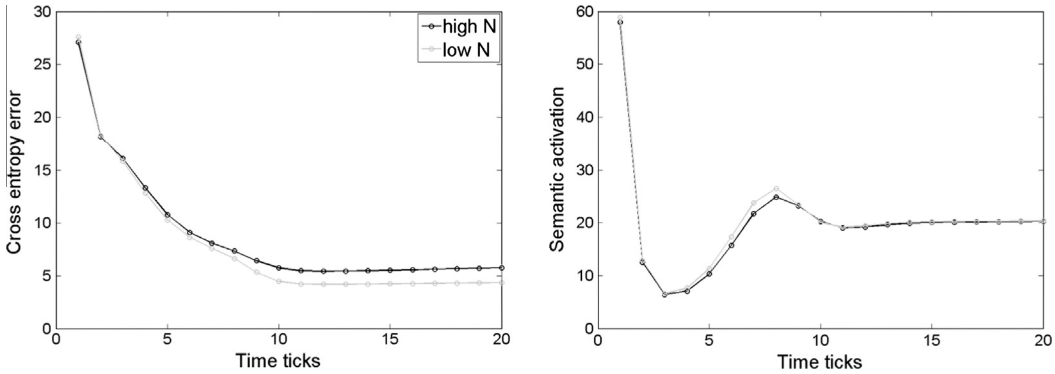

Fig. 8. Influences of orthographic neighborhood size (*N*) on network error (left) and activation (right).

of variance with orthographic neighborhood size (large vs. small) as a between-items variable and time (ticks 1–20) as a within-items variable.

Error was significantly higher for words with many as compared to few orthographic neighbors, $F(1,138) = 5.02$, $p < .05$, $\eta^2 = .04$. This orthographic neighborhood effect interacted with time, $F(19,2622) = 3.35$, $p < .05$. Planned comparisons revealed that the orthographic neighborhood effect was significant at ticks 8–20 (all *F*s > 3.94, all *p*s < .05).

There was no main effect of orthographic neighborhood size on semantic activation, $F(1,138) = 2.33$, $p = .13$. However, orthographic neighborhood size interacted with time, $F(19,2622) = 4.37$, $p < .01$, with significantly higher activation for words with a smaller orthographic neighborhood at ticks 1, 2, and 4–8 (all *F*s > 7.65, *ps* < .01, $\eta^2 > .05$).

3.7.3. Discussion

Network error was higher for words with a greater number of orthographic neighbors, in line with larger N400 amplitudes for these words (Holcomb et al., 2002; Laszlo & Federmeier, 2011). On the other hand, semantic activation was slightly lower for words with many neighbors, which is opposite to the larger N400 amplitudes.

It seems somewhat difficult to relate network measures to influences of orthographic neighborhood size on lexical decision performance for a couple of reasons. First, the empirical evidence is inconsistent. Second, orthographic neighborhood size presumably primarily influences lexical decisions via decision criteria concerning orthographic familiarity, and has been successfully simulated at the orthographic level of representation without assuming influences from semantics (Grainger & Jacobs, 1996). Therefore, potential influences of orthographic neighborhood size on semantic processes as simulated in the present model may simply not be relevant to decision criteria and hence outcomes in lexical decisions (see Seidenberg & McClelland, 1989).

4. General discussion

The present research investigated the computational mechanisms underlying the N400 component of the ERP by relating N400 amplitude modulations to variations in two measures taken from an attractor network model of word meaning, semantic network error and total semantic activation. Our goal was to investigate potential cognitive processes underlying N400 amplitude modulations, rather than their neural realization. We simulated a number of N400 effects obtained in human empirical research on word processing: influences of semantic priming (Simulation 1), semantic richness (Simulation 2), word frequency (Simulation 3), and word repetition (Simulation 4), as well as influences of semantic richness on repetition effects (Simulation 5), word frequency on repetition effects (Simulation 6), and orthographic neighborhood size (Simulation 7). A summary of the results is presented in Table 1.

In all seven simulations, semantic network error was in the same direction as N400 amplitudes. Like the N400, error was lower for semantically related target words (see Fig. 2; Bentin et al., 1985), higher for words with richer semantic representations (Fig. 3; Holcomb et al., 1999; Kounios & Holcomb, 1994; Laszlo & Federmeier, 2011; Müller et al., 2010; Rabovsky et al., 2012a, 2012c) and higher for low frequency words (Fig. 4; Van Petten & Kutas, 1990). Furthermore, error was lower for repeated words (Fig. 5; Nagy & Rugg, 1989), and this repetition-induced decrease was stronger for words with richer semantic representations (Fig. 6; Kounios & Holcomb, 1994; Rabovsky et al., 2012b), and for low frequency words (Fig. 7; Rugg, 1990). Finally, error was higher for words with many orthographic neighbors (Fig. 8; Holcomb et al., 2002; Laszlo & Federmeier, 2011).

In contrast, there was less consistency between semantic activation and the N400. Like N400 amplitudes, activation was larger for words with richer semantic representations (Fig. 3). However, activation also increased with frequency, repetition, semantic priming, and a smaller orthographic neighborhood (Figs. 2 and 4–8), which is opposite to N400 results. Instead, the results may be more in line with the notion that increased semantic activation can facilitate lexical decisions (see Section 4.3 for further discussion). Importantly, our results suggest an interesting relation between N400 amplitudes and semantic network error.

4.1. The N400 as implicit prediction error in semantic memory

On a psychological level, network error has been conceptualized as implicit prediction error, with model-generated and correct output representing implicit prediction and observation, respectively (Elman, 1990; McClelland, 1994; O'Reilly et al., 2012; Rogers & McClelland, 2008; see Section 1.3). Thus, the consistent relation between N400 amplitudes and error values in our semantic network model suggests that N400 amplitudes reflect implicit prediction error in semantic memory (McClelland, 1994).

As discussed in Section 1, the model provides a simplified description of word recognition processes (including only abstract word form units and semantic features), but as the present work focuses on understanding the mechanisms underlying the N400 ERP component which is seen as a functionally specific electrophysiological indicator of semantic processing (Kutas & Federmeier, 2011), it seems appropriate to use a model that focuses on semantic computations. The model illustrates several important points. We assume that N400 amplitudes reflect implicit prediction error at the level of semantic features. Note that we do not claim that the semantic features used in our model precisely mirror those in the human conceptual system; rather, they provide a window into those representations (see McRae et al., 2005, for a more detailed discussion). This implicit prediction error is presumably influenced by both general and conditional probabilities of occurrence. That is, it is influenced by the general occurrence probability of any individual feature, as well as by the probability of each feature to occur together with the other features in the currently relevant set (correlated features). Implicit prediction error is influenced further by the mapping between the input (word form) and the semantic features, that is, by the probability of the semantic features given the specific input units. Based on the empirical evidence (Kutas & Federmeier, 2011), we assume that N400 amplitudes reflect implicit semantic prediction error independent of input modality and type.[3] Therefore, prediction error at the semantic layer is key, and semantic computations could be initiated by other forms of input (e.g., a picture). There is no specific relation to lexical representations which are assumed to modulate prediction error at the semantic layer via the strength of their connections with semantic features, not qualitatively different from other forms of input representation. Even though the present model does not include components such as attention or intention, we also assume, based on empirical evidence, that the processes underlying N400 amplitudes are (like many other processes) modifiable by attention (McCarthy & Nobre, 1993) and strategy (Lau, Holcomb, & Kuperberg, 2013), as well as the inherent salience of certain features (Paczynski & Kuperberg, 2012). Thus, the implicit predictive processes presumably occuring constantly and automatically can be enhanced by attention and active prediction, resulting in stronger modulations of the N400 (Lau et al., 2013).

The suggestion that N400 amplitudes reflect implicit semantic prediction error seems consistent with proposals relating other ERP negativities to prediction errors in other domains. These include the more frontally distributed feedback-related negativity (FRN) that has been related to reward prediction error (Cavanagh, Frank, Klein, & Allen, 2010; Chase et al., 2011; Cohen & Ranganath, 2007; Holroyd & Coles, 2002; Walsh & Anderson, 2012), as well as the mismatch negativity (MMN), which is presumably related to implicit prediction errors in processing auditory stimulus sequences (Garrido et al., 2009; Wacongne et al., 2012), and the visual mismatch negativity (vMMN), its visual counterpart (Kimura, Kondo, Ohira, & Schroger, 2011; Winkler & Czigler, 2012).

In line with such an account in terms of implicit prediction error, N400 amplitudes seem to depend crucially on the similarity between actual observations and implicit anticipations based on represented occurrence probabilities as extracted from previously experienced regularities. From this perspective, N400 amplitudes are larger for semantic violations and low cloze probability sentence continuations, because low or zero cloze words occur less frequently in the respective contexts and are therefore less expected. The represented occurrence probability for low frequency words is presumably generally rather low, resulting in enhanced implicit prediction error and thus enhanced N400 amplitudes for low frequency words. Recent exposure to a word presumably enhances its represented occurrence probability, giving rise to repetition-induced reductions of implicit prediction error and hence N400 amplitudes. Increased N400 amplitudes for concepts with a greater number of semantic features can be explained by assuming that for every semantic feature, it is on average less probable that it is involved in the currently relevant concept than that it is not involved because semantic representations are sparse. This is definitely true for the features in the norms of McRae et al. (2005), and hence in the present model, where on average each of the 2526 features occurs in only 2.87 out of the 541 concepts, so that the average probability for a feature to be included in the currently relevant concept is below 1%. Thus, the presence of any specific semantic feature is improbable, and even though the norms obviously do not cover the entire space of concepts and features, it seems reasonable to assume that this pattern generalizes beyond this reduced semantic space. Hence, every semantic feature may signal implicit prediction error when it is (overall unexpectedly) involved in the current concept, resulting in higher cumulative implicit prediction error and thus larger N400 amplitudes for words with more semantic features.

From the present perspective, larger N400 amplitudes for larger orthographic neighborhoods (e.g. Holcomb et al., 2002; Laszlo & Federmeier, 2011) may be explained by the fact that the conditional probability for specific semantic features is strongly dependent on the orthographic input, or in other words, the mapping is less unequivocal and more difficult for words with many orthographic neighbors. That is, if the visual input *dish* is presented, the conditional probability for the semantic features of the orthographic neighbors *fish* and *wish* is presumably enhanced, whereas the conditional probability for the features of *dish* may be not as high as in case when there are no orthographically similar words (because the input *ish* is not a clear predictor of the semantic features of *dish*). Thus, implicit prediction error might be higher

[3] Even though, as noted above, there are slight differences in spatial distributions across the scalp (see e.g., Van Petten & Luka, 2006).

for words with many orthographic neighbors both because implicit expectations for semantic features that are not part of the target semantic representation are misleadingly increased, and because the implicit expectations for the semantic features that are part of the target semantic representation are lower than they would be with a more unique word form.

The influence of orthographic neighborhood size on the N400 has been shown to be independent of lexical type, so that N400 amplitudes were larger for words and pseudowords than for acronyms and illegal letter strings (Laszlo & Federmeier, 2011). Even though we did not compare these types of stimuli, the orthographic neighborhood size simulation seems relevant, and Laszlo and Federmeier's findings appear to fit well with the idea that N400 amplitudes reflect implicit semantic prediction error. Specifically, misleadingly high implicit expectations for the semantic features of the orthographic neighbors of both words and pseudowords, and somewhat weakened expectations for the correct semantic features of the words (due to the lack of unambiguity of the predictions derived from visual input) could both contribute to higher implicit prediction error for these types of stimuli. On the other hand, acronyms (i.e., lexical stimuli without orthographic neighbors) presumably elicit expectations only for the correct semantic features (because the input pattern is quite unique), and illegal strings presumably correctly elicit very little expectation for semantic features at all, so that implicit prediction error would be low in both cases.

Interestingly, Laszlo and Federmeier (2011) also report an influence of neighbor frequency on the N400, with larger N400 amplitudes for words with higher as compared to lower frequency neighbors. Although we did not simulate this study, it is plausible that error induced by orthographic neighbors would be stronger if these neighbors were of higher frequency. This is because stronger connections for the higher frequency neighbors than for the target word would aggravate the misleadingly high expectation of the semantic features of the orthographic neighbor and the lower expectation of the correct semantic features of the target.

Semantic priming effects, that is, smaller N400 amplitudes to target words following a semantically related as compared to an unrelated prime word, can be due to two mechanisms. First, related words may often co-occur in language, so that the occurrence of one word may increase the probability of encountering the other word, and therefore implicit prediction error would be lower (similar to the sentence context manipulations discussed above). This is presumably the case for some types of associated words, namely those that tend to follow one another in language (e.g., *dog-bark* or *bread-butter*). Word co-ocurrence was not implemented in the present model's training regime. Second, related words may share semantic features. The present semantic priming simulation used featurally-similar words (concepts). From the present perspective, one can assume that the current brain state (including semantic activation) is mostly a good predictor of the next brain state (including semantic activation), because usually the environment (and its meaning) does not all of a sudden change completely. Thus, less change (e.g. due to overlapping semantic features between subsequent stimuli) entails lower prediction error (see McRae, de Sa, & Seidenberg, 1997, for further discussion).

The present model-based account of the N400 in terms of implicit prediction error in the semantic system conceptually overlaps with some of the verbally described theories on the N400. Specifically, the current account is consistent with the views that the N400 reflects the ease or difficulty of accessing features in semantic memory (Kutas & Federmeier, 2011; Van Berkum, 2009), or the match or mismatch between expected and encountered semantic features (Paczynski & Kuperberg, 2012).

Debruille (2007) suggested that N400 amplitudes reflect semantic inhibition. As erroneously predicted semantic features may need to be inhibited for the network to settle into the correct semantic pattern, inhibition might be expected to substantially co-vary with implicit semantic prediction error. It may be interesting to prospectively explore the extent to which inhibitory activity co-varies with error values, and whether and how inhibition and error can be disentangled.

Lau et al. (2008) suggest that N400 effects reflect facilitated lexical access. The authors discuss predictive processes as a source of facilitation, so that there exists substantial conceptual overlap with the current account. However, their focus is on facilitation at the lexical level, whereas the current account suggests that the N400 reflects implicit prediction error at the level of semantics which can in principle be activated from different stimulus modalities and domains.

It has also been suggested that the N400 reflects semantic unification (Baggio & Hagoort, 2011), that is, the integration of semantic information accessed from the current stimulus with the current semantic context, a process that gives rise to an overall semantic representation that is not pre-stored in memory. In our view, the effort involved in such a unification process crucially depends on the internally represented probability for the semantic features of the current stimulus to co-occur with those already activated by context, so that semantic unification depends on implicit semantic prediction and prediction error. From our perspective, the present account is more general than the semantic unification theory in that it relates N400 amplitudes not only to the probability of occurrence of semantic features insofar as it can be deduced from the currently active set, but to the probability of occurrence of semantic features more generally (which for example can be additionally deduced from general occurrence probability, i.e., frequency).

As outlined above, some theories posit the N400 at specific levels in a processing hierarchy (lexical access, higher-level contextual integration) and have problems accommodating N400 modulations induced by manipulations affecting a different level (e.g., an integration account has problems accommodating word frequency effects; see Kutas & Federmeier, 2011, for further discussion). In contrast, understanding N400 amplitudes as reflecting implicit prediction error in the semantic system can naturally explain N400 modulations induced by variables affecting different levels of processing. This is the case as long as the manipulation influences the match (at the level of semantic features) between implicit anticipations (based on previously experienced regularities across levels of representation as represented in semantic memory) and the current stimulus. However, we are not arguing that the N400 reflects these different levels. The N400 is assumed to reflect implicit prediction error in semantic memory, and variables affecting different levels of processing can influence N400 amplitudes insofar as they modulate the local context and thereby modulate conditional probabilities in semantics. From this perspective, contextual manipulations (sentences, categories, events) that modulate the conditional occurrence probability for specific aspects of meaning should modulate N400 amplitudes.

It is crucial to note that the sort of implicit prediction discussed here is not meant to correspond to active, conscious, explicit prediction of specific lexical items. Indeed, violations of active predictions of specific lexical items may not be reflected in N400 amplitudes, but instead in subsequent ERP positivities (Federmeier, Wlotko, De Ochoa-Dewald, & Kutas, 2007; Lau et al., 2013; Paczynski & Kuperberg, 2011; Van Petten & Luka, 2012). One possibility might be that N400 amplitudes reflect implicit prediction error and adaptation in long-term semantic memory, whereas subsequent positivities may reflect a similar process – prediction error and update - in active working memory (e.g., Polich, 2007). Kuperberg (2013) more specifically suggested that the basis for this prediction error and update in working memory may be the combinatorial integration of semantic features with other types of representations. She argues that the posterior late positivity is triggered by a prediction error for a specific semantic-syntactic mapping that is disconfirmed by the input, whereas the anterior late positivity is triggered by a prediction error for a specific semantic-wordform mapping (specific lexical item) that is disconfirmed by the input.

Finally, we would like to note that although N400 amplitudes were not consistently related to the amount of activation in the semantic layer in our simulations, if expected semantic features are pre-activated prior to the occurrence of the relevant stimulus, then transient semantic activation (i.e., during the computations required for moving from the initial/predicted activation state to the correct activation state) can be correlated with semantic error. Thus, in that sense, there is a correspondence between Laszlo and Plaut's (2012) model and the current model.

4.2. The N400 and implicit memory formation

Error-driven learning is common in connectionist modeling, reflecting the assumption that implicit prediction error drives implicit memory formation. The assumption that humans generate implicit anticipations based on an internal model of the environment, and adapt their internal model based on the discrepancy between prediction and observation, is widely shared in cognitive science and neuroscience (Clark, 2013; den Ouden, Daunizeau, Roiser, Friston, & Stephan, 2010; den Ouden, Friston, Daw, McIntosh, & Stephan, 2009; Friston, 2005, 2009; McClelland, 1994; McLaren, 1989; Schultz & Dickinson, 2000; Tobler, O'Doherty, Dolan, & Schultz, 2006). This conception has also been related to language processing in a recent review, where Kutas et al. (2011, p. 201) discuss the view that "an unexpected item triggers updating in a learning signal, where probability likelihoods are being adjusted for the future", with comprehenders benefitting "by gaining an accurate model of their linguistic environment."

Intriguingly, if N400 amplitudes reflect semantic network error, and this error drives connection adaptation, then enhanced N400 amplitudes should imply enhanced connection adaptation, that is, enhanced implicit memory formation. Although there is a large literature on N400 effects related to memory retrieval (i.e. the so-called 'familiarity' N400; see e.g., Curran, 2000; Voss & Federmeier, 2011), there is not a great deal of evidence on the relation between N400 amplitudes and memory encoding. However, a few studies may be relevant.

Schott, Richardson-Klavehn, Heinze, and Düzel (2002) reported larger N400 amplitudes during a learning phase to predict implicit memory during test (as assessed by repetition priming effects on stem completion in the absence of explicit memory). In addition, a recognition memory study by Meyer, Mecklinger, and Friederici (2007) seems pertinent. The authors combined a study phase that included coherent and anomalous sentences with a recognition memory test phase. Words presented as semantically anomalous sentence endings (which as usual elicited larger N400 amplitudes) produced the strongest N400 old/new effect in the test phase (i.e., more centro-parietal positivity for repeated as compared to new words). In addition, Meyer et al. reported significant correlations across participants between overall N400 amplitudes during study and N400 old/new effects during test. Interestingly, even though they discussed the N400 old/new effect in the context of familiarity-based recognition, its modulation was not accompanied by influences on explicit recognition memory performance so that it may have reflected implicit repetition priming. Relatedly, Olichney et al. (2000) observed impaired recognition and recall performance but intact N400 incongruity and repetition effects in amnesic patients, in line with preserved implicit memory formation in amnesia.

Furthermore, a longitudinal study by Friedrich and Friederici (2006) showed that age-adequate as compared to poor language skills at 30 months were predicted by the presence (vs. absence) of N400 effects at 19 months. Friedrich and Friederici (2010) found that 12 month old infants with high early word production showed an N400 semantic priming effect whereas those with low early word production did not, even for words that they supposedly understood (as rated by their parents). Friedrich and Friederici (2010, p. 70) concluded that "although the combined results [. . .] suggest that the early functioning of the N400 neural mechanisms strongly interacts with the children's language development, we still cannot describe this interaction in detail." They discuss several interpretations of the observed relation, including the possibility "that the mechanisms underlying the N400 elicitation are directly involved in the word learning process." Interestingly, this possibility would be naturally implied by assuming that N400 amplitudes reflect implicit semantic prediction error that drives learning and adaptation in the lexical-semantic system.

In summary, although further evidence is required, there are indeed some hints suggesting a relation between N400 amplitudes and implicit memory formation, in line with the proposed model-based account of N400 amplitudes as reflecting implicit prediction error in the semantic system. Because implicit prediction error is assumed to drive implicit memory formation, yielding an intrinsic relation between error and adaptation, it is not entirely clear whether N400 amplitudes reflect implicit prediction error, or whether they might rather reflect the connection adaptations driven by implicit prediction error.

4.3. Semantic activation and lexical decision performance

Our simulations yielded little support for the notion that N400 amplitudes covary directly with the amount of semantic activation. Although activation was larger for words with richer semantic representations, which is in line with N400 amplitudes (Amsel, 2011; Holcomb et al., 1999; Kounios & Holcomb, 1994; Laszlo & Federmeier, 2011; Müller et al., 2010; Rabovsky et al., 2012a, 2012c; West & Holcomb, 2000), activation also increased with semantic priming, frequency, repetition, and a small orthographic neighborhood, which is opposite to N400 results (Bentin et al., 1985; Holcomb et al., 2002; Laszlo & Federmeier, 2011; Nagy & Rugg, 1989; Van Petten & Kutas, 1990). Instead, if anything, our simulations may be taken to suggest a relation between semantic activation and behavioral performance with high semantic activation facilitating lexical decisions. It is important to note that our discussion here focuses on lexical decisions only rather than behavioral performance in general, as influences of semantic activation can depend on the type of task (e.g., Chen & Mirman, 2012). Assuming that high semantic activation can facilitate lexical decisions, the model successfully predicted facilitation for semantically related target words (McRae & Boisvert, 1998; Neely, 1991), repeated words (Scarborough et al., 1977), high frequency words (Forster & Chambers, 1973), and words with richer semantic representations (Pexman et al., 2008). In addition, it predicts enhanced repetition-induced facilitation for low frequency words (Forster & Davis, 1984), and for words with richer semantic representations (Rabovsky et al., 2012b).

It has been suggested that semantic activation can drive behavioral performance, presumably by feeding into decision processes in lexical decision tasks (Grondin et al., 2009). This could be realized in, for example, random walk models of decision making (Joordens, Piercey, & Azerbehi, 2003; Ratcliff, Gomez, & McKoon, 2004). That is, high semantic activation may facilitate deciding that a stimulus is a word and not a nonword because words have meaning whereas nonwords do not. Thus, high activation levels in the semantic system may facilitate lexical decisions because decision thresholds between words and nonwords are at least partly based on semantic activation and thus may be easier to cross when semantic activation is high. In line with such a view, Laszlo and Plaut (2012) also modeled lexical decisions based on thresholded semantic activation.

As more extensively discussed in Section 1, however, semantic activation as simulated in the present model should be related to lexical decision performance only in cases in which decision criteria based on semantic activation support optimal (i.e., as fast as possible while maintaining acceptable accuracy) lexical decisions. For instance, when simulating orthographic neighborhood effects, semantic activation may not be a good predictor of lexical decision performance because orthographic neighborhood size supposedly has its impact via decision criteria at the orthographic level of representation (Grainger & Jacobs, 1996).

5. Conclusion

Using a feature-based attractor network model of word meaning to simulate seven N400 effects obtained in empirical research on word processing, we consistently found a close correspondence between network error and N400 amplitudes. Based on conceptualizing network error as implicit prediction error (Elman, 1990; McClelland, 1994; O'Reilly et al., 2012; Rogers & McClelland, 2008), these results suggest that N400 amplitudes reflect implicit prediction error in semantic memory (McClelland, 1994).

Acknowledgements

This work was supported by a doctoral scholarship from the German National Academic Exchange Service (DAAD) to Milena Rabovsky and Natural Sciences and Engineering Council Grant OGP0155704 and National Institutes of Health Grant HD053136 to Ken McRae. We thank Daniel Schad for valuable advice and discussion, and David Plaut, Gina Kuperberg, and an anonymous reviewer for insightful and constructive comments on earlier versions of the manuscript.

References

Amsel, B. D. (2011). Tracking real-time neural activation of conceptual knowledge using single-trial event-related potentials. *Neuropsychologia, 49*(5), 970–983. http://dx.doi.org/10.1016/j.neuropsychologia.2011.01.003.

Andrews, S. (1989). Frequency and neighborhood effects on lexical access – Activation or search. *Journal of Experimental Psychology-Learning Memory and Cognition, 15*(5), 802–814. http://dx.doi.org/10.1037/0278-7393.15.5.802.

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review, 4*(4), 439–461. http://dx.doi.org/10.3758/Bf03214334.

Baggio, G., & Hagoort, P. (2011). The balance between memory and unification in semantics: A dynamic account of the N400. *Language and Cognitive Processes, 26*(9), 1338–1367. http://dx.doi.org/10.1080/01690965.2010.542671.

Bar, M. (2009). The proactive brain: Memory for predictions. *Philosophical Transactions of the Royal Society B-Biological Sciences, 364*(1521), 1235–1243. http://dx.doi.org/10.1098/rstb.2008.0310.

Barber, H., Vergara, M., & Carreiras, M. (2004). Syllable-frequency effects in visual word recognition: Evidence from ERPs. *Neuroreport, 15*(3), 545–548. http://dx.doi.org/10.1097/01.wnr.0000111325.38420.80.

Barrett, S. E., & Rugg, M. D. (1989). Event-related potentials and the semantic matching of faces. *Neuropsychologia, 27*(7), 913–922.

Barrett, S. E., & Rugg, M. D. (1990). Event-related potentials and the semantic matching of pictures. *Brain and Cognition, 14*(2), 201–212. http://dx.doi.org/10.1016/0278-2626(90)90029-N.

Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society B—Biological Sciences, 364*(1521), 1281–1289. http://dx.doi.org/10.1098/rstb.2008.0319.

Bentin, S., Mccarthy, G., & Wood, C. C. (1985). Event-related potentials, lexical decision and semantic priming. *Electroencephalography and Clinical Neurophysiology, 60*(4), 343–355.

Blackford, T., Holcomb, P. J., Grainger, J., & Kuperberg, G. R. (2012). A funny thing happened on the way to articulation: N400 attenuation despite behavioral interference in picture naming. *Cognition, 123*(1), 84–99. http://dx.doi.org/10.1016/j.cognition.2011.12.007.

Brown, C., & Hagoort, P. (1993). The processing nature of the N400: Evidence from masked priming. *Journal of Cognitive Neuroscience, 5*(1), 34–44. http://dx.doi.org/10.1162/jocn.1993.5.1.34.

Cavanagh, J. F., Frank, M. J., Klein, T. J., & Allen, J. J. (2010). Frontal theta links prediction errors to behavioral adaptation in reinforcement learning. *Neuroimage, 49*(4), 3198–3209. http://dx.doi.org/10.1016/j.neuroimage.2009.11.080.

Chase, H. W., Swainson, R., Durham, L., Benham, L., & Cools, R. (2011). Feedback-related negativity codes prediction error but not behavioral adjustment during probabilistic reversal learning. *Journal of Cognitive Neuroscience, 23*(4), 936–946.

Chen, Q., & Mirman, D. (2012). Competition and cooperation among similar representations: Toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological Review, 119*(2), 417–430.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181–204. http://dx.doi.org/10.1017/S0140525X12000477.

Cohen, M. X., & Ranganath, C. (2007). Reinforcement learning signals predict future decisions. *Journal of Neuroscience, 27*(2), 371–378. http://dx.doi.org/10.1523/JNEUROSCI.4421-06.2007.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.). *Attention and performance* (Vol. 6, pp. 535–555). New York: Academic Press.

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*(4), 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*(3), 371–414.

Curran, T. (2000). Brain potentials of recollection and familiarity. *Memory & Cognition, 28*(6), 923–938. http://dx.doi.org/10.3758/Bf03209340.

Debruille, J. B. (2007). The N400 potential could index a semantic inhibition. *Brain Research Reviews, 56*(2), 472–477. http://dx.doi.org/10.1016/j.brainresrev.2007.10.001.

den Ouden, H. E. M., Daunizeau, J., Roiser, J., Friston, K. J., & Stephan, K. E. (2010). Striatal prediction error modulates cortical coupling. *Journal of Neuroscience, 30*(9), 3210–3219. http://dx.doi.org/10.1523/Jneurosci.4458-09.2010.

den Ouden, H. E. M., Friston, K. J., Daw, N. D., McIntosh, A. R., & Stephan, K. E. (2009). A dual role for prediction error in associative learning. *Cerebral Cortex, 19*(5), 1175–1185. http://dx.doi.org/10.1093/cercor/bhn161.

Doya, K., Ishii, S., Pouget, A., & Rao, R. P. N. (Eds.). (2007). *Bayesian brain: Probabilistic approaches to neural coding*. MIT Press.

Dunabeitia, J. A., Aviles, A., & Carreiras, M. (2008). NoA's ark: Influence of the number of associates in visual word recognition. *Psychonomic Bulletin & Review, 15*(6), 1072–1077. http://dx.doi.org/10.3758/Pbr.15.6.1072.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*(2), 179–211. http://dx.doi.org/10.1016/0364-0213(90)90002-E.

Federmeier, K. D., & Laszlo, S. (2009). Time for meaning: Electrophysiology provides insights into the dynamics of representation and processing in semantic memory. *Psychology of Learning and Motivation: Advances in Research and Theory, 51*(51), 1–44. http://dx.doi.org/10.1016/S0079-7421(09)51001-8.

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research, 1146*, 75–84. http://dx.doi.org/10.1016/j.brainres.2006.06.101.

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior, 12*(6), 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*(4), 680–698.

Friedrich, M., & Friederici, A. D. (2006). Early N400 development and later language acquisition. *Psychophysiology, 43*(1), 1–12. http://dx.doi.org/10.1111/j.1469-8986.2006.00381.x.

Friedrich, M., & Friederici, A. D. (2010). Maturing brain mechanisms and developing behavioral language skills. *Brain and Language, 114*(2), 66–71. http://dx.doi.org/10.1016/j.bandl.2009.07.004.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B—Biological Sciences, 360*(1456), 815–836. http://dx.doi.org/10.1098/rstb.2005.1622.

Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences, 13*(7), 293–301. http://dx.doi.org/10.1016/j.tics.2009.04.005.

Furl, N., van Rijsbergen, N. J., Treves, A., Friston, K. J., & Dolan, R. J. (2007). Experience-dependent coding of facial expression in superior temporal sulcus. *Proceedings of the National Academy of Sciences of the United States of America, 104*(33), 13485–13489. http://dx.doi.org/10.1073/pnas.0702548104.

Garrido, M. I., Kilner, J. M., Stephan, K. E., & Friston, K. J. (2009). The mismatch negativity: A review of underlying mechanisms. *Clinical Neurophysiology, 120*(3), 453–463. http://dx.doi.org/10.1016/j.clinph.2008.11.029.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review, 103*(3), 518–565. http://dx.doi.org/10.1037/0033-295x.103.3.518.

Grainger, J., Oregan, J. K., Jacobs, A. M., & Segui, J. (1989). On the role of competing word units in visual word recognition – The neighborhood frequency effect. *Perception & Psychophysics, 45*(3), 189–195. http://dx.doi.org/10.3758/Bf03210696.

Grondin, R., Lupker, S. J., & McRae, K. (2009). Shared features dominate semantic richness effects for concrete concepts. *Journal of Memory and Language, 60*(1), 1–19. http://dx.doi.org/10.1016/j.jml.2008.09.001.

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*(6), 938–950.

Holcomb, P. J., Kounios, J., Anderson, J. E., & West, W. C. (1999). Dual-coding, context-availability, and concreteness effects in sentence comprehension: An electrophysiological investigation. *Journal of Experimental Psychology: Learning Memory and Cognition, 25*(3), 721–742.

Holroyd, C. B., & Coles, M. G. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review, 109*(4), 679–709.

Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychonomic Bulletin & Review, 10*(4), 785–813.

Joordens, S., Piercey, C. D., & Azerbehi, R. (2003). From word recognition to lexical decision: A random walk along the road of harmony. In D. D. F. Detje, & H. Schaub (Eds.), *Proceedings of the fifth international conference on cognitive modeling* (pp. 141–146).

Kiefer, M. (2005). Repetition-priming modulates category-related effects on event-related potentials: Further evidence for multiple cortical semantic systems. *Journal of Cognitive Neuroscience, 17*(2), 199–211.

Kimura, M., Kondo, H., Ohira, H., & Schroger, E. (2011). Unintentional temporal context-based prediction of emotional faces: An electrophysiological study. *Cerebral Cortex*. http://dx.doi.org/10.1093/cercor/bhr244.

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*(4), 804–823.

Kuperberg, G. R. (2013). The proactive comprehender: What event-related potentials tell us about the dynamics of reading comprehension. In B. Miller, L. Cutting, & P. McCardle (Eds.), *Unraveling reading comprehension: Behavioral, neurobiological, and genetic components* (pp. 176–192). Baltimore: Paul Brookes Publishing.

Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. In M. Bar (Ed.), *Predictions in the brain: Using our past to generate a future* (pp. 190–207). New York: Oxford University Press.

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

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential

(ERP). *Annual Review of Psychology, 62*, 621–647. 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*(4427), 203–205.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307*(5947), 161–163.

Kutas, M., Neville, H. J., & Holcomb, P. J. (1987). A preliminary comparison of the N400 response to semantic anomalies during reading, listening and signing. *Electroencephalography and Clinical Neurophysiology Supplement, 39*, 325–330.

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*(3), 271–281. http://dx.doi.org/10.1016/j.bandl.2011.09.001.

Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 effects of prediction from association in single-word contexts. *Journal of Cognitive Neuroscience, 25*(3), 484–502.

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (De)constructing the N400. *Nature Reviews Neuroscience, 9*(12), 920–933. http://dx.doi.org/10.1038/nrn2532.

Lucas, M. (2000). Semantic priming without association: A meta-analytic review. *Psychonomic Bulletin & Review, 7*(4), 618–630.

McCarthy, G., & Nobre, A. C. (1993). Modulation of semantic processing by spatial selective attention. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section, 88*(3), 210–219.

McClelland, J. L. (1994). The interaction of nature and nurture in development: A parallel distributed processing perspective. In P. E. P. Bertelson & G. d'Ydewalle (Eds.). *International perspectives on psychological science* (Vol. 1). United Kingdom: Erlbaum.

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science, 1*(1), 11–38. http://dx.doi.org/10.1111/j.1756-8765.2008.01003.x.

McLaren, I. (1989). The computational unit as an assembly of neurons: An implementation of an error correcting learning algorithm. In R. Durbin, C. Miall, & G. Mitchison, & G. (Eds.), *The computing neuron* (pp. 160–178). Amsterdam, the Netherlands: Addison-Wesley.

McRae, K., & Boisvert, S. (1998). Automatic semantic similarity priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(3), 558–572.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, 37*(4), 547–559.

McRae, K., Cree, G. S., Westmacott, R., & de Sa, V. R. (1999). Further evidence for feature correlations in semantic memory. *Canadian Journal of Experimental Psychology, 53*(4), 360–373.

McRae, K., de Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General, 126*, 99–130.

Meyer, P., Mecklinger, A., & Friederici, A. D. (2007). Bridging the gap between the semantic N400 and the early old/new memory effect. *Neuroreport, 18*(10), 1009–1013.

Mirman, D., & Magnuson, J. S. (2008). Attractor dynamics and semantic neighborhood density: Processing is slowed by near neighbors and speeded by distant neighbors. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(1), 65–79. http://dx.doi.org/10.1037/0278-7393.34.1.65.

Müller, O., Dunabeitia, J. A., & Carreiras, M. (2010). Orthographic and associative neighborhood density effects: What is shared, what is different? *Psychophysiology, 47*(3), 455–466. http://dx.doi.org/10.1111/j.1469-8986.2009.00960.x.

Nagy, M. E., & Rugg, M. D. (1989). Modulation of event-related potentials by word repetition: The effects of inter-item lag. *Psychophysiology, 26*(4), 431–436.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theory. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition*. Hillsdayle, NJ: Erlbaum.

Niedeggen, M., Rösler, F., & Jost, K. (1999). Processing of incongruous mental calculation problems: Evidence for an arithmetic N400 effect. *Psychophysiology, 36*(3), 307–324.

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*(3), 507–518.

O'Connor, C. M., Cree, G. S., & McRae, K. (2009). Conceptual hierarchies in a flat attractor network: Dynamics of learning and computations. *Cognitive Science, 33*(4), 665–708. http://dx.doi.org/10.1111/j.1551-6709.2009.01024.x.

Olichney, J. M., Van Petten, C., Paller, K. A., Salmon, D. P., Iragui, V. J., & Kutas, M. (2000). Word repetition in amnesia – Electrophysiological measures of impaired and spared memory. *Brain, 123*, 1948–1963. http://dx.doi.org/10.1093/brain/123.9.1948.

O'Reilly, R. C., Munakata, Y., Frank, M. J., Hazy, T. E., & Contributors (2012). *Computational cognitive neuroscience. Wiki book* (1 ed.). <http://ccnbook.colorado.edu>.

Ostwald, D., Spitzer, B., Guggenmos, M., Schmidt, T. T., Kiebel, S. J., & Blankenburg, F. (2012). Evidence for neural encoding of Bayesian surprise in human somatosensation. *Neuroimage, 62*(1), 177–188. http://dx.doi.org/10.1016/j.neuroimage.2012.04.050.

Paczynski, M., & Kuperberg, G. R. (2011). Electrophysiological evidence for use of the animacy hierarchy, but not thematic role assignment, during verb-argument processing. *Language and Cognitive Processes, 26*(9), 1402–1456. http://dx.doi.org/10.1080/01690965.2011.580143.

Paczynski, M., & Kuperberg, G. R. (2012). Multiple influences of semantic memory on sentence processing: Distinct effects of semantic relatedness on violations of real-world event/state knowledge and animacy selection restrictions. *Journal of Memory and Language, 67*(4), 426–448. http://dx.doi.org/10.1016/j.jml.2012.07.003.

Pearlmutter, B. A. (1995). Gradient calculation for dynamic recurrent neural networks: A survey. *IEE Transactions on Neural Networks, 6*, 1212–1228.

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*(1), 161–167. http://dx.doi.org/10.3758/Pbr.15.1.161.

Pexman, P. M., Holyk, G. G., & Monfils, M. H. (2003). Number-of-features effects and semantic processing. *Memory & Cognition, 31*(6), 842–855.

Pexman, P. M., Lupker, S. J., & Hino, Y. (2002). The impact of feedback semantics in visual word recognition: Number-of-features effects in lexical decision and naming tasks. *Psychonomic Bulletin & Review, 9*(3), 542–549.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103*(1), 56–115. http://dx.doi.org/10.1037/0033-295x.103.1.56.

Plunkett, K., & Elman, J. L. (1997). *Exercises in rethinking innateness: A handbook for connectionist simulations*. Cambridge MA: MIT Press.

Polich, J. (2007). Updating p300: An integrative theory of P3a and P3b. *Clinical Neurophysiology, 118*(10), 2128–2148. http://dx.doi.org/10.1016/j.clinph.2007.04.019.

Rabovsky, M., Alvarez, C. J., Hohlfeld, A., & Sommer, W. (2008). Is lexical access autonomous? Evidence from combining overlapping tasks with recording event-related brain potentials. *Brain Research, 1222*, 156–165. http://dx.doi.org/10.1016/j.brainres.2008.05.066.

Rabovsky, M., Sommer, W., & Abdel Rahman, R. (2012a). Depth of conceptual knowledge modulates visual processes during word reading. *Journal of Cognitive Neuroscience, 24*(4), 990–1005.

Rabovsky, M., Sommer, W., & Abdel Rahman, R. (2012b). Implicit word learning benefits from semantic richness: Electrophysiological and behavioral evidence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(4), 1076–1083. http://dx.doi.org/10.1037/a0025646.

Rabovsky, M., Sommer, W., & Abdel Rahman, R. (2012c). The time course of semantic richness effects in visual word recognition. *Frontiers in Human Neuroscience, 6*. http://dx.doi.org/10.3389/fnhum.2012.00011.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*(1), 79–87.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review, 111*(1), 159–182. http://dx.doi.org/10.1037/0033-295x.111.1.159.

Rogers, T. T., & McClelland, J. L. (2008). Precis of semantic cognition: A parallel distributed processing approach. *Behavioral and Brain Sciences, 31*(6), 689–749. http://dx.doi.org/10.1017/S0140525x0800589x.

Rugg, M. D. (1990). Event-related brain potentials dissociate repetition effects of high-frequency and low-frequency words. *Memory & Cognition, 18*(4), 367–379.

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), 1–17.

Schott, B., Richardson-Klavehn, A., Heinze, H. J., & Düzel, E. (2002). Perceptual priming versus explicit memory: Dissociable neural correlates at encoding. *Journal of Cognitive Neuroscience, 14*(4), 578–592.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275*(5306), 1593–1599.

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual Review of Neuroscience, 23*, 473–500.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96*(4), 523–568.

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*(2), 260–273. http://dx.doi.org/10.1016/j.cogbrainres.2005.02.006.

Stark, C. E. L., & McClelland, J. L. (2000). Repetition priming of words, pseudowords, and nonwords. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(4), 945–972. http://dx.doi.org/10.1037//0278-7393.26.4.945.

Summerfield, C., Egner, T., Greene, M., Koechlin, E., Mangels, J., & Hirsch, J. (2006). Predictive codes for forthcoming perception in the frontal cortex. *Science, 314*(5803), 1311–1314. http://dx.doi.org/10.1126/science.1132028.

Tobler, P. N., O'Doherty, J. P., Dolan, R. J., & Schultz, W. (2006). Human neural learning depends on reward prediction errors in the blocking paradigm. *Journal of Neurophysiology, 95*(1), 301–310.

Tyler, L. K., & Moss, H. E. (2001). Towards a distributed account of conceptual knowledge. *Trends in Cognitive Sciences, 5*(6), 244–252. http://dx.doi.org/10.1016/S1364-6613(00)01651-X.

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

Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word-frequency in event-related brain potentials. *Memory & Cognition, 18*(4), 380–393.

Van Petten, C., & Luka, B. J. (2006). Neural localization of semantic context effects in electromagnetic and hemodynamic studies. *Brain and Language, 97*(3), 279–293. http://dx.doi.org/10.1016/j.bandl.2005.11.003.

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology, 83*(2), 176–190. http://dx.doi.org/10.1016/j.ijpsycho.2011.09.015.

Van Petten, C., & Rheinfelder, H. (1995). Conceptual relationships between spoken words and environmental sounds: Event-related brain potential measures. *Neuropsychologia, 33*(4), 485–508.

Voss, J. L., & Federmeier, K. D. (2011). FN400 potentials are functionally identical to N400 potentials and reflect semantic processing during recognition testing. *Psychophysiology, 48*(4), 532–546. http://dx.doi.org/10.1111/j.1469-8986.2010.01085.x.

Wacongne, C., Changeux, J. P., & Dehaene, S. (2012). A neuronal model of predictive coding accounting for the mismatch negativity. *Journal of Neuroscience, 32*(11), 3665–3678. http://dx.doi.org/10.1523/Jneurosci.5003-11.2012.

Walsh, M. M., & Anderson, J. R. (2012). Learning from experience: Event-related potential correlates of reward processing, neural adaptation, and behavioral choice. *Neuroscience & Biobehavioral Reviews*. http://dx.doi.org/10.1016/j.neubiorev.2012.05.008.

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*(6), 1024–1037.

Winkler, I., & Czigler, I. (2012). Evidence from auditory and visual event-related potential (ERP) studies of deviance detection (MMN and vMMN) linking predictive coding theories and perceptual object representations. *International Journal of Psychophysiology, 83*(2), 132–143. http://dx.doi.org/10.1016/j.ijpsycho.2011.10.001.

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*(6), 664–676.

Ziegler, J. C., & Perry, C. (1998). No more problems in Coltheart's neighborhood: Resolving neighborhood conflicts in the lexical decision task. *Cognition, 68*(2), B53–B62.