Brain & Language 120 (2012) 271–281

Contents lists available at SciVerse ScienceDirect

Brain & Language

journal homepage: www.elsevier.com/locate/b&l

A neurally plausible Parallel Distributed Processing model of Event-Related Potential word reading data

Sarah Laszlo [a,*], David C. Plaut [b,c]

[a] Department of Psychology, State University of New York, Binghamton, NY, United States
[b] Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, United States
[c] Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA, United States

ARTICLE INFO

Article history:
Accepted 1 September 2011
Available online 25 September 2011

Keywords:
Computational modeling
Parallel Distributed Processing
Event-Related Potentials
N400
Visual word recognition

ABSTRACT

The Parallel Distributed Processing (PDP) framework has significant potential for producing models of cognitive tasks that approximate how the brain performs the same tasks. To date, however, there has been relatively little contact between PDP modeling and data from cognitive neuroscience. In an attempt to advance the relationship between explicit, computational models and physiological data collected during the performance of cognitive tasks, we developed a PDP model of visual word recognition which simulates key results from the ERP reading literature, while simultaneously being able to successfully perform lexical decision—a benchmark task for reading models. Simulations reveal that the model's success depends on the implementation of several neurally plausible features in its architecture which are sufficiently domain-general to be relevant to cognitive modeling more generally.

1. Introduction

Comprehending meaning from text—visual word recognition—is a pervasive and fundamental cognitive process that is studied by researchers using a wide variety of methodologies. In broad strokes, cognitive scientists seek to characterize the component processes involved, cognitive neuroscientists seek to map those processes onto neural signatures, and computational modelers seek to make explicit the interactions that occur between the representations involved. Each of these methodologies has strengths that can supplement the weaknesses of others, and often important discoveries are made when two or more of them are combined—for example, when psychophysiology provides a time course for proposed cognitive processes or when a computational model shows that a particular cognitive architecture can in fact produce the pattern of results it has been formulated to explain.

Interplay between cognitive science and computational modeling in the domain of visual word recognition has involved the parallel development of two prominent but very different modeling frameworks: one utilizing learned representations and a uniform set of computational principles—the Parallel Distributed Processing (PDP) approach (e.g., Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989)—and another which de-emphasizes learning and relies on different types of computations in different functional pathways—the so-called "dual-route" or "dual-process" approach (e.g., Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Perry, Ziegler, & Zorzi, 2007). Each of these approaches has its own strengths and weaknesses, but in aggregate both of them are highly successful in simulating a number of results from behavior and neuropsychology. For example, one compilation of effects that recent models have been successful in simulating (Perry et al., 2007), includes 13 items, from diverse tasks such as lexical decision, reading aloud, and many variants of priming, as well as several items pertaining to performance in dyslexia. However, there is one area in which even the most sophisticated of current models is lacking, as agreed upon by proponents of both the PDP and dual-process frameworks (e.g., Harm & Seidenberg, 2004; Perry et al., 2007), as well as advocates of other modeling techniques in other domains (e.g., Bayesian modeling; see Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010). That area is contact with data from cognitive neuroscience and neurophysiology. It is widely hoped that more contact with cognitive neuroscience can provide constraining data on appropriate internal dynamics for models, and that more contact with data from neuroscience can improve the neural plausibility of models largely based on behavior.

Interestingly, this need for more contact with cognitive neuroscience in computational investigations of visual word recognition has coincided with a need for more contact with computational models in similar investigations conducted using the Event-Related Potential (ERP) methodology. It has begun to be commonly noted that theories about the representations and computations involved in reading stemming from ERP data have become specific enough that it would be desirable to test them by instantiating

* Corresponding author. Address: Department of Psychology, Binghamton University, 4400 Vestal Parkway East, Binghamton, NY 13902, United States.
E-mail address: cogneuro@alum.mit.edu (S. Laszlo).

0093-934X/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.bandl.2011.09.001

them as computational models (e.g., Barber & Kutas, 2007; van Berkum, 2008). For example, a recent series of ERP studies pertaining to the "obligatory semantics" view of visual word recognition has presented data cast as strongly consonant with PDP models, while less supportive of dual-process models (Laszlo & Federmeier, 2007, 2008, 2009, 2011). These studies have focused on the N400 ERP component, which, as discussed in more detail below, is thought to be a functionally specific marker of attempted semantic access (see Kutas & Federmeier, 2011, for review). It has now been shown several times that even meaningless items with little resemblance to lexically represented items can engage the semantic access thought to be indexed by the N400, both in sentences (Laszlo & Federmeier, 2009) and in unconnected streams of text (Laszlo, Stites, & Federmeier, in press)—that is, an attempt to access semantics appears to be obligatory for all orthographic inputs, even consonant strings like XFQ. Further, the N400s elicited by meaningless illegal strings respond to manipulation of lexical characteristics such as orthographic neighborhood size (i.e., Coltheart's N, the number of words that can be created by changing one letter of a target item; Coltheart, Davelaar, Jonasson, & Besner, 1977) and neighbor frequency in a manner both quantitatively and qualitatively similar to that demonstrated by words (Laszlo & Federmeier, 2011). These data have been taken as supportive of PDP models in that they seem to reveal a language processing system which (1) does not require an item to have a lexical representation, or even be similar to an item with a lexical representation, in order to make some contact with semantics and (2) performs what appear to be indistinguishable computations on different input types, regardless of factors like lexicality or the regularity/consistency of spelling-sound correspondences. Further, the degree to which an attempt at semantic access occurs for meaningless items appears to be strongly related to their similarity to items with associated semantics (i.e., words, acronyms), a result which is consonant with the fact that the distributed representations preferred by PDP models tend to associate similar inputs with similar outputs, to a degree determined by the amount of overlap between representations.

In contrast, the ERP results seem to be less supportive of dual-process models, insofar as such models include lexical mediation between orthographic input and semantics (e.g., Perry et al., 2007), making it difficult or impossible for items such as consonant strings, which are neither lexically represented nor similar to items that are, to contact semantics. Note that a lexically mediated system could potentially be made to allow illegal strings contact with semantics by lowering the threshold of lexical activation that needs to be met in order for semantics to be activated. That is, the many lexical entries that overlap slightly with illegal strings could be activated weakly, and the aggregation of this weak activity over many units could be allowed to be passed forward to semantics. However, such a system is no longer strongly lexicalized, in that the internal representations that mediate between orthography and semantics are now essentially distributed—that is, many units participate in the representation of each input, and the strength of activation in those units is proportional to the degree of overlap with the input. This will be true not just for nonwords but also for words as, of course, words overlap with other words to differing degrees.

Another potential mismatch between the ERP results pertaining to meaningless, illegal strings and dual-process models occurs because of one of the core properties of dual-process models: orthographic inputs tend to differentially engage separable processing streams depending on the regularity of their spelling-sound correspondence. This characteristic seems incongruent with the repeated finding that items with irregular spelling-sound correspondences (acronyms, consonant strings), elicit waveforms that are qualitatively and quantitatively quite similar to those elicited by items with regular spelling sound correspondence (words, pseudowords) up to and including the N400 portion of the ERP (Laszlo & Federmeier, 2007).

The fact that these ERP data have been explicitly cast as supportive of one particular theoretical framework invites an attempt to test the obligatory semantics view by trying to simulate key ERP data in a PDP model of the type they are claimed to support. An attempt to test the obligatory semantics view by instantiating its assumptions in an explicit computational model would be useful not only in advancing a theoretical position present in the ERP literature—it would also provide new information about the degree to which the internal dynamics of a reading model constructed with PDP principles match the internal dynamics of the groups of neurons that are actually performing the task in the brain. Currently, there is limited constraint on the internal dynamics of cognitive reading models, as they are all based almost entirely on behavioral data, which is fundamentally *end state* data. That is, while strong inferences about internal processing can and have been made on the basis of, for example, RT or naming latency data, these data do not provide *direct* evidence about the processes occurring between when an item is presented and when a response is made—only data about the final consequences of those processes. ERPs, in contrast, can be collected continuously between when an item is presented and when a response is made, and can, in fact, be collected even when no overt response is made. Further, ERPs can be divided into well-specified *components*, which have been robustly replicated as reflecting particular cognitive functions.

The N400 component, for a particularly relevant example, is strongly tied with attempted semantic access. The designation of the N400 as a semantic component is based on a variety of converging results, including its functional properties, its neural generators, and the functional anatomy of components which precede it. The N400 is known to respond to a wide variety of semantic manipulations such as congruity with sentence and discourse context (Kutas & Hillyard, 1984; van Berkum, Hagoort, & Brown, 1999), semantic association (Nobre & McCarthy, 1995), and item concreteness (Kounios & Holcomb, 1994), to name only a few, while not being sensitive to other types of linguistic manipulations, such as those of syntactic constraint (Kutas & Hillyard, 1983), or font size (Kutas & Hillyard, 1980). Converging evidence from intracranial EEG (Nobre & McCarthy, 1995), MEG (Halgren et al., 2002), and the Event-Related Optical Signal (EROS; Tse et al., 2007), as well as patterns of diminished N400 in brain damage (Hagoort, Brown, & Swaab, 1996) all point to a primary source of the N400 in the left anterior temporal lobe, a region strongly linked with semantic processing (e.g., McCarthy, Nobre, Bentin, & Spencer, 1995; Nobre & McCarthy, 1995). Finally, the N400 has been argued to occur not only in the correct brain areas, but also in the correct temporal window, to subserve semantic access, based on both the neural generators and functional properties of the sensory and form-based components that precede it (see Grainger & Holcomb, 2009, for extensive review). In sum, the functional specificity of the N400 component is a particularly useful property for model-building, as its clear link with semantic processing permits a direct comparison with semantic representations and processes in a model.

The goal of the present work is to test the assumptions of the obligatory semantics view of N400 processing in a PDP model that continuously simulates N400 amplitude. Three particular considerations are of importance. First, can such a model produce N400-like dynamics at all—that is, can we produce a PDP model the semantic activation of which resembles the morphology of the N400 component? To our knowledge there are no other implemented computational models of N400 processing, so this is not assured. We chose to link N400 amplitude with amount of activation in

the semantic layer of representation in our model, on the basis of the N400's strong link in the literature with semantic access (as just discussed) as well as findings that, at least in the context of reading unconnected text, N400 amplitude represents the number of semantic features being activated in response to a particular input. (e.g., Laszlo & Federmeier, 2007, 2011). Larger (more negative) N400s are elicited by items which might be expected to activate more semantic features, such as items higher in concreteness (e.g., West & Holcomb, 2000), or items with larger orthographic neighborhood sizes. The morphology of the N400 is well-characterized as essentially a curve which rises monotonically to a single peak, and then decreases monotonically throughout the remainder of its time course—Fig. 1 displays several N400 potentials representative of those we sought to simulate. To be successful, the mean amount of activity in the model's semantic layer must develop similarly, without, for example, additional oscillations. In this fashion, the model is constrained not only to reach some end state in a manner consistent with the data (as is the case in behavioral models), but also to perform in a manner consistent with the data throughout its evolution over time.

The second consideration is: will the dynamics of the semantic layer in the model further mirror critical results supportive of the obligatory semantics view? In seeking to answer this question, we chose to focus our simulations on data from the single-item ERP corpus (Laszlo & Federmeier, 2011), as it is both uniquely appropriate for computational modeling and also representative of the key data in support of obligatory semantics. The availability of single-item ERPs enables items analysis (e.g., items multiple regression), in addition to the more typical parametric analysis available from ERP reading studies. This makes the single-item ERP corpus a particularly appropriate target for computational modeling, as it is advantageous to model items effects, not just item aggregated, factorial effects, whenever possible. For the model to be successful, in addition to showing the broad characteristics of the N400, it must also produce simulated N400s that are consistent with the critical findings from the single-item ERP corpus (described in detail below).

Finally, it is important that the model also be able to perform the behavioral task of lexical decision, as lexical decision is among the most common benchmark tasks for computational reading models. Literate adults, though they do not receive extensive training on performing lexical decisions while learning to read, are able to make them quite easily in a lab setting. In imitation of this situation, the model is never explicitly trained on lexical decision but is asked to make lexical decisions on the basis of a simple thresholding procedure after training is complete. Attempting to implement this additional capability in the model helps to ensure that, insofar as it is able to simulate results previous models do not—from the domain of ERPs—it is also able to simulate the fundamental behavioral data that decades of visual word recognition modeling have been built on. Without this additional ability, the ERP model would not truly be tied to its thematic predecessors (e.g., Harm & Seidenberg, 2004; Plaut et al., 1996; Seidenberg & McClelland, 1989), which would be unfortunate given the significant insights those models have provided into visual word recognition. Simulating both electrophysiological and behavioral data is a more challenging task than simulating the ERP data alone, but a worthwhile one: it lays a foundation for a much more complete, holistic model than ignoring the behavioral data would. Further, challenging the model to perform lexical decision instantiates an incremental approach to computational modeling (Perry et al., 2007) by extending a preliminary ERP model that focused on the ERP data alone (Laszlo & Plaut, 2011). A criterion for model success was that, by the end of processing each input, the model be able to produce a signal that could reliably differentiate meaningful items (words and acronyms) from non-meaningful items (pseudowords and illegal strings).

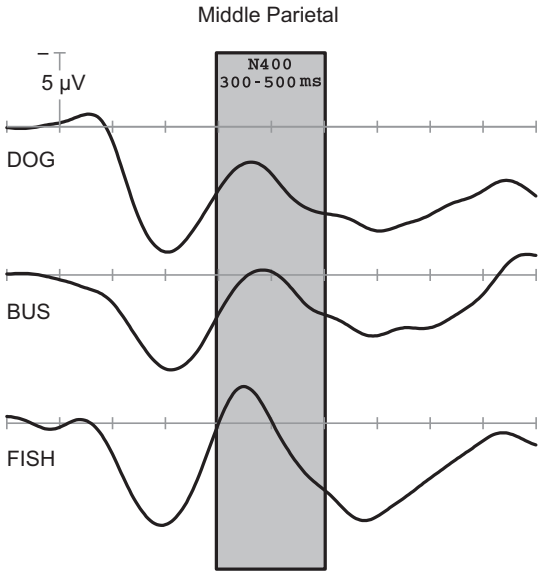

Fig. 1. Representative single-item ERPs averaged over 120 participants, but not over items. The middle parietal electrode site, where N400 effects are most prominent, is displayed. Typical N400 morphology is visible in the 300–500 ms N400 window (boxed), for the words DOG, BUS, and FISH. In this figure, as in all ERP figures, negative is plotted up.

In developing a model of ERP data, we considered it critical to incorporate some of the most general properties of the neurons which produce the ERP signal. The vast majority of the brain-generated electrical potential measured at the scalp is produced by the synchronous firing of excitatory and inhibitory post-synaptic potentials by cortical neurons arranged in an open-field configuration (see Fabiani, Gratton, & Federmeier, 2007, for review). Thus, we departed from previous reading models by trying to handle excitation and inhibition in the model in a manner more true to what is understood about the neural configuration of excitation and inhibition (see, e.g., Crick & Asanuma, 1986). This was accomplished in three ways. First, we separated excitation and inhibition in the model, such that individual units could have excitatory outgoing projections or inhibitory outgoing projections, but never both, as is true of cortical neurons. This arrangement can be observed in Fig. 2, which presents a schematic of the ERP model. Second, we limited the distribution of inhibitory connections, such that they could occur only within, but never between, levels of representation in the model. This decision was motivated by the fact that connections between cortical areas are largely excitatory, with inhibitory connections occurring largely within a given cortical area. This feature of the model is also visible in Fig. 2. Finally, we severely limited the number of inhibitory units in the model—each excitatory layer has only a single associated inhibitory unit—in accordance with the finding that the large majority of neurons in the cortex are excitatory (e.g., White, 1989). Each of these neurally plausible adjustments to the way excitation and inhibition are handled in the model represent a departure from previous reading models (e.g., Harm & Seidenberg, 2004), in that inhibition is typically unconstrained in such models, with individual units able to have both positive and negative outgoing connections, inhibitory connections allowed between levels of representation, and, because excitation and inhibition are not separated, essentially equal numbers of excitatory and inhibitory units.

In what follows, we first present the relevant phenomena from the single-item ERP corpus in some detail, in order to directly motivate the simulations that follow. Then, in two simulations, we explore a number of questions pertaining to the ability of a PDP system to successfully simulate the ERP data. First, we attempt

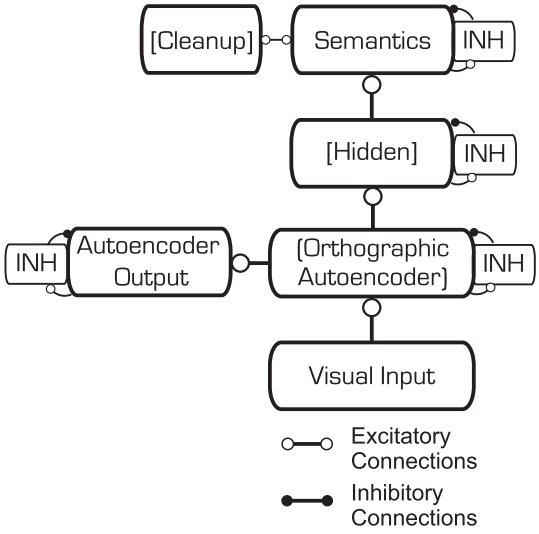

Fig. 2. Schematic of the ERP model. Lines with empty circles indicate excitatory connections, lines with filled circles indicate inhibitory connections. INH stands for "inhibitory," and each INH bank consists of only 1 unit. Note that no units have both excitatory and inhibitory outgoing connections, and that inhibition is always within, never between, levels of representation.

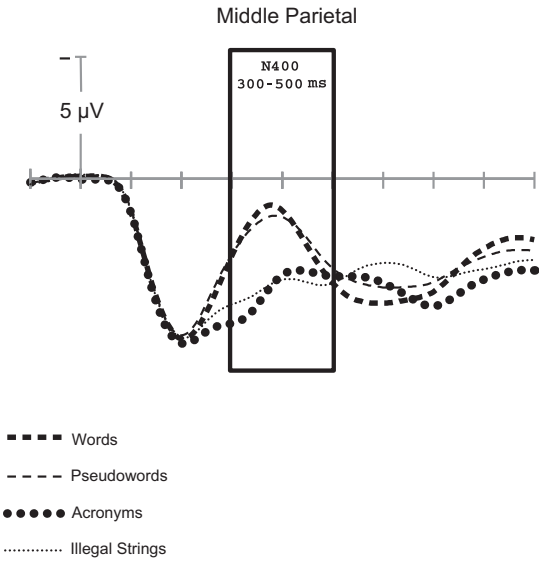

Fig. 3. Orthographic neighborhood size effect in item-aggregated ERPs. Item types with high N (words, pseudowords) elicited larger N400s than item types with lower N (acronyms, illegal strings).

to determine whether a PDP system can produce internal dynamics which resemble ERP morphology at all. If this is accomplished, we then seek to determine whether such a system can produce the results thought to be supportive of the obligatory semantics view of N400 processing: namely a strong effect of orthographic neighborhood size which acts similarly for lexical and non-lexical items. Importantly, if the model is able to correctly simulate the key ERP findings, its ability to perform lexical decisions is assessed as an additional metric of success. Finally, the contribution of the separation of excitation and inhibition in the model to the model's ability to simulate the ERP data is examined.

2. Target phenomena: Event-Related Potentials

A detailed report of the methods and results of the single-item ERP corpus is available elsewhere (Laszlo & Federmeier, 2011). However, for clarity, we describe here the nature of that data set and the key results that will act as target phenomena for the simulations presented below. 120 participants in the single-item ERP study viewed an unconnected list of words (e.g., HAT), pseudowords (e.g., KOF), acronyms (e.g., DVD), and meaningless illegal strings (e.g., NHK), while monitoring the stream for English proper names (e.g., SARA, DAVE). No response was required for the critical item types, in order to keep the critical ERPs free from response related components. This task, as well as the item types presented, replicated Laszlo and Federmeier (2007). Acronyms were backsorted on the basis of a post-test such that only acronym items that individual participants were familiar with were included in that participant's averaged waveforms. Event-Related Potentials were formed by averaging at each of the scalp electrodes time-locked to the onset of each of the critical items. In the case of single-item ERPs, averaging was done over participants only, not over items. More typical, item-aggregated ERPs (representing, for example, the response to all words) were formed by averaging over both items and participants.

One of the most striking findings in the single-item data is that individual lexical characteristics (e.g., orthographic neighborhood size, neighbor frequency), tend to be much stronger predictors of N400 amplitude than lexical type (e.g., word or pseudoword). This is demonstrated in Fig. 3, in the case of orthographic neighborhood size. As is evident in Fig. 3, items with high N (words, pseudowords) elicit larger N400s than items with low N (acronyms, illegal strings), and this is true regardless of lexicality. That is, though pseudowords are presumably not semantically represented, they elicit similar N400s to words, because of their similarity on N—the same is true when comparing acronyms and illegal strings. This can be quantified as a main effect of N on N400 mean amplitude, but no effect of lexicality and no interaction between the two (see Laszlo & Federmeier, 2011, for details of statistical analysis).

The second critical finding we consider in the simulations below is that, at an items level, the slopes relating N400 mean amplitude to orthographic neighborhood size are qualitatively and quantitatively quite similar for lexical and non-lexical items—this is, of course, reflected as the lack of interaction between N and lexicality in the factorial analysis. This result is visible in Fig. 4 (reproduced

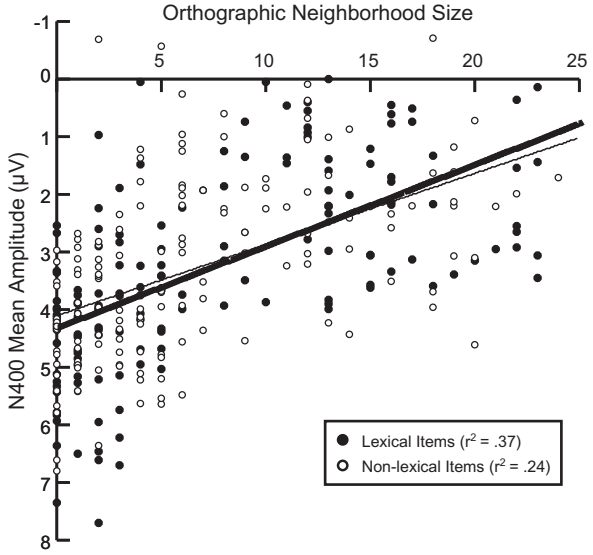

Fig. 4. Orthographic neighborhood size effect in single item ERPs. N400 mean amplitude is computed over the middle parietal electrode site in the 300–500 ms post stimulus onset epoch. Lexical items (words and acronyms) are represented by filled dots, non-lexical items (pseudowords and illegal strings) are represented by empty dots. Note that the slopes representing the relationship between orthographic neighborhood size and N400 mean amplitude are quite similar. Reproduced from Laszlo and Federmeier (2011).

from Laszlo & Federmeier, 2011), which displays a scatter plot of items N400 mean amplitude vs. orthographic neighborhood size along with single regression trendlines for lexical and non-lexical items. Note that the distributions of N400 mean amplitudes for lexical and non-lexical items are almost completely overlapping, as are the trendlines depicting the relationship between orthographic neighborhood size and N400 mean amplitude for the two lexical types. Thus, the N effect is quite similar for lexical and non-lexical items. In addition, automated stepwise regression analysis of the single-item ERP corpus revealed that N is, by far, the strongest predictor of unique N400 variance of those lexical variables considered (length, N, neighbor frequency, number of lexical associates, and frequency of top associate were all considered in Laszlo & Federmeier, 2011; subsequent analysis has extended the list to include bigram frequency, concreteness, imageability, number of senses, and noun verb ambiguity; Laszlo, unpublished data). The prominence of the N effect, combined with other findings indicating that, unlike effects of variables such as concreteness or written frequency, it is maintained both with repetition and in sentence context (Laszlo & Federmeier, 2007, 2008, 2009), altogether make it particularly relevant for simulations exploring the obligatory semantics view of the N400.

3. Simulation 1

3.1. Methods

The architecture of the ERP model is depicted in Fig. 2. A 15-unit visual input layer represents the visual features of each of three letters in five non-overlapping slots. The visual input layer feeds into a 20-unit orthographic autoencoder, which was pre-trained to reproduce the visual input on a copy of the 15 input units. The autoencoder feeds through a 50-unit hidden layer to a 50-unit semantic layer with an associated 30-unit semantic cleanup layer. At the semantic layer, relatively sparse, arbitrary semantic representations were trained to be associated with the visual inputs, in accordance with the fact that, for morphologically simple words in English, orthography-semantics mappings are largely arbitrary. Semantic targets consisted of random bit patterns over the 50 semantic units—that is, semantic features were not learned but were arbitrarily assigned, with the constraint that each unit be active in at least one semantic target. Either 3 or 7 features were active in semantics for each target. The numbers 3 and 7 were chosen simply so that semantic representations would be fairly sparse (i.e., 6% of features active for a representation with three features, 14% active for a representation with seven features). Two different numbers of features were chosen so that effects of semantic concreteness could be explored in future versions of the model using the same materials: the N400 is known to be sensitive to semantic concreteness (Kounios & Holcomb, 1994). Weights on connections between levels of representation were constrained to be positive-only. Each layer of representation (except for the cleanup and input layers) has one associated inhibitory unit, connected as depicted in Fig. 2.

For excitatory units, the standard logistic (sigmoid) function was used to compute unit activations. For the inhibitory units, a multi-linear activation function was used, with a slope of 1 from inputs of zero through an inflection point, and a slope of 2 from the inflection point onward (see Fig. 5). The multi-linear activation function was used in order to approximate the presence in the brain of separate populations of inhibitory neurons with varying temporal response properties—that is, the fact that some inhibitory neurons respond more quickly with stimulation than others (e.g., Benado, 1994; Traub, Miles, & Wong, 1989). As is visible in Fig. 5, the multi-linear activation function is formally identical to the sum of (1) a linear activation function that begins immediately with even small amounts of input and (2) an identical linear activation function that does not begin until some threshold of activation is passed (the "elbow"). Because it takes time for activation to build up in the network, the result is that the steeper portion of the inhibitory function is not used until later in network time than the shallower portion. In this way, even though the network only has one inhibitory unit at each level of representation, that one unit is able to approximate the function of separate units with different temporal properties. The inhibitory activation function is unbounded—allowing the single inhibitory unit associated with each level of representation to produce significant inhibition—and the location of the inflection point in activation space for each inhibitory unit is a fixed parameter in the model. Output units (i.e., units in the semantic layer or the orthographic output layer in the autoencoder) are additionally constrained such that their activation decays towards zero as the inverse square root of their raw, logistic activation. Thus, units that are strongly activated tend to stay strongly activated, while units that are weakly activated tend to decay towards zero activation. This procedure is reminiscent of a k-winners-take-all function (O'Reilly, 1996a), in that it allows only the units with the strongest activations to remain active, and quiets all the rest, but differs in that the number of units that are able to remain active is dynamic.

Training was accomplished by back-propagating cross-entropy error through time (Hinton, 1989; Rumelhart, Hinton, & Williams,

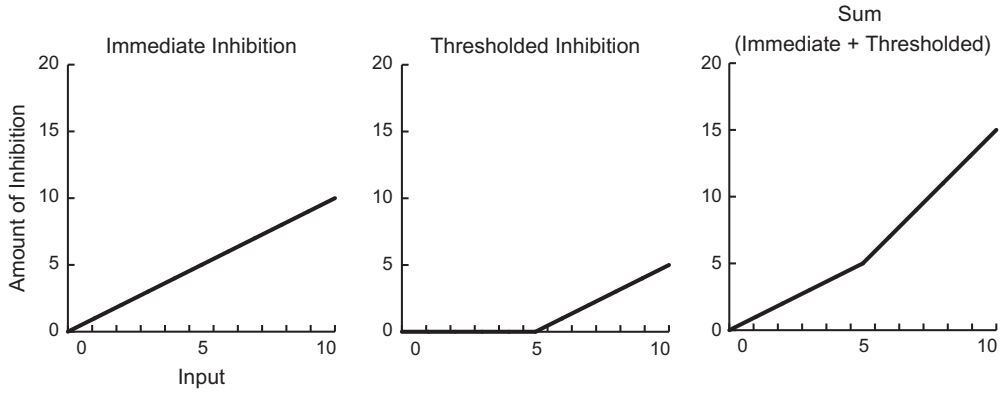

Fig. 5. The sum of an immediate, linear inhibition function (left) and a delayed, linear inhibition function (center) is a multilinear ("elbowed") function (right). In the model, inhibition is a function of input activation, not time, so its relationship to truly time dependent inhibition is only approximate. The slopes displayed here are the actual slopes used in the simulation (i.e., a slope of 1 to the inflection point, and a slope of 2 after). The inflection point in the model's inhibition function is a fixed parameter. Both "Amount of Inhibition" and "Time" are in arbitrary units.

1986). Additional constraints were added to the back-propagation procedure to assure that excitatory weights were always positive and inhibitory weights were always negative. First, the minimum outgoing weight of excitatory units is a fixed parameter in the model such that in the present implementation of the back-propagation algorithm, no weight change is made that would cause a weight to be smaller than its fixed minimum. Second, inhibitory weights were fixed to random, negative values at the beginning of training and were not updated subsequently. Thus, it was impossible for connections designated as excitatory to have negative values, or for connections designated as inhibitory to have positive values.

In order to keep the scale of the model small, there are only 10 letters in its vocabulary: seven consonants (SNCBDPT) and three vowels (OIU). Of the possible 1000 strings of letters that could be formed with 10 letters in three slots (10^3), we designated 62 as "words" and 15 as "acronyms." Words were constrained to have a CVC structure, and acronyms could have any letters in the 1st or 3rd position, but were constrained to have a consonant in the 2nd position—this was done to create a structural difference between the representations of words and acronyms and also to ensure that the orthographic neighborhood sizes of acronyms would be smaller than that of words, as is true in the single-item ERP corpus. Within the limited vocabulary of the model, words had a within-set N of 6.83, and acronyms had a within-set N of 0.8.

Before semantic training commenced, the autoencoder was trained to reproduce the orthography of each of the 77 semantically represented (i.e., word and acronym) items (see the Section 3.2). This was done to ensure that, even before semantic training began, the network had some knowledge about the orthographic structure of input items. By forcing the network to condense, and reconstruct, orthographic representations prior to the onset of semantic training, the autoencoder ensures that orthographic structure will be emphasized in subsequent processing. The model learns, during autoencoder training, that inputs with interior consonants are dispreferred, through the simple fact that more words—items with interior vowels—are presented than acronyms—items with interior consonants. This information is important to the model, as without it illegal strings tend to produce too much semantic activation, by virtue of their structural similarity with acronyms. A related consequence of pre-training particular orthographic structures is that acronyms form strong internal representations in the model, without which they would be unable to activate semantics sufficiently because they are dissimilar to and less frequent than words. In essence, what the model learns by pre-training on orthographic structure it is that internal consonants are dispreferred, and thus should generally not pass much activation forward, except in the specific cases with robust representations in the autoencoder—that is, except for acronyms.

After autoencoder training was complete, the semantic training phase began, during which time the network was trained to activate the correct (although arbitrary) semantic features for each of the 77 semantically represented items, while simultaneously being trained to keep all features in semantics "off" for a large set of "wordlike" nonwords. The wordlike nonwords consisted of the 1155 (77 items $*$ 15 input features) items that could be formed by flipping one bit in the input representations of the 77 semantically represented items. That is, by changing a single one in an input representation to a zero, or vice versa. Although there were more wordlike nonwords than semantically represented items in the training corpus, each word was presented to the network during training 50 times more frequently than each nonword. One way to think about training on wordlike nonwords is that it approximates training the network to not link semantics with "mistakes," much like training a learning reader that a word misspelled by one letter is not the same as the word itself.

On each training trial, the visual input for one of the items in the training corpus was clamped on, and activation was allowed to propagate through the network for 12 time steps with no accumulation of error. Targets continued to be presented for a subsequent four time steps, during which time error was accumulated. At the end of 16 time steps, the trial ended, the network was reset to its initial state, and the next trial began. Words and acronyms were 50 times more likely to be selected as the input for each trial than wordlike nonwords. A single training epoch consisted of 1232 (77 + 1155) trials, however not every item was necessarily trained in each epoch as words and acronyms were more likely to be selected than nonwords (e.g., a single word could be selected 50 times, meaning that not every item would be selected in every 1232 trial epoch). After 9000 epochs of training in this fashion, the network was tested on 441 items: the 62 words and 15 acronyms it was trained on, in addition to 279 illegal strings (nonwords with central consonants) and 85 pseudowords (nonwords with central vowels) to which the network was not exposed during training. The target for all illegal strings and pseudowords was for all semantic units to remain off.

3.2. Results

3.2.1. Autoencoder

The orthographic autoencoder was trained to reproduce the visual inputs corresponding to the 62 words and 15 acronyms on a copy of the input units. For the autoencoder, as for subsequent analyses pertaining to the model's semantic performance, an output is considered correct if the Euclidean distance between the output representation produced for an item and the target representation for that item is lower than the distance between the output and the target for any other item. After 3000 epochs of training, the autoencoder's performance was perfect (100%).

3.2.2. Semantics

After 9000 epochs of training, the network was 93% (411/441 items) accurate in producing either correct semantics (in the case of words and acronyms) or silence in the semantic layer (in the case of pseudowords and illegal strings). Of the 30 errors, 15 occurred for pseudowords, and 15 occurred for illegal strings—all items that were actually trained (words and acronyms) were correctly linked with semantics. Fig. 6 displays the mean activation in semantics over time for words, acronyms, pseudowords and illegal strings—that is, the data corresponding to the item aggregated ERPs displayed in Fig. 3. Two important features of the data are visible in Fig. 6.

First, by the end of the processing epoch, the model has successfully separated words and acronyms from pseudowords and illegal strings, meaning that a simple threshold on mean semantic activation is sufficient for separating semantically represented items from non-represented items in 90% percent of cases. That is, the model can accurately make lexical decisions based on a single, set activation threshold, despite having never been explicitly trained on lexical decision. In particular, with an activation threshold of 0.0579, 100% of words and acronyms are correctly accepted, and 88% (319/364) of pseudowords and illegal strings are correctly rejected.

Second, as in the N400 data, words and pseudowords tend to elicit more activity in semantics than do acronyms and illegal strings. To investigate the relationships between N, lexicality, and mean semantic activation in the model, we conducted a simultaneous multiple regression on mean semantic activation with N, lexicality, and the N $\times$ lexicality interaction as predictors. Mean semantic activation for an item in the model was computed as the average amount of activation elicited by that item across all 16 time steps. This analysis revealed that, just as in the ERPs, there

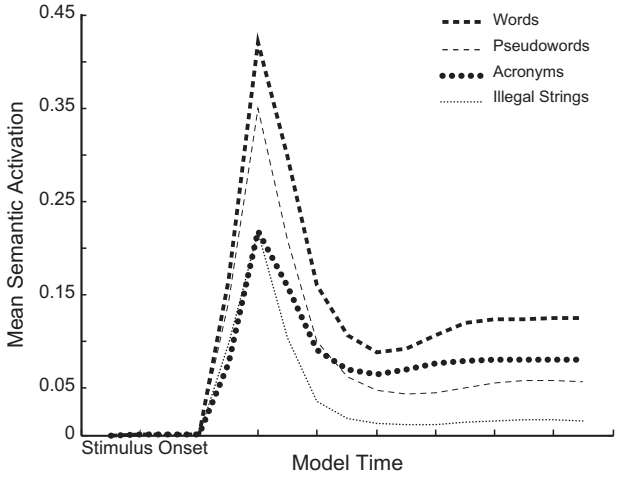

Fig. 6. Orthographic neighborhood size effect in item-aggregated model output. Item types with high N (words, pseudowords) elicit larger simulated N400s than item types with lower N (acronyms, illegal strings). Notice also that, by the end of the epoch, semantically represented items (words, acronyms) are separated from non-represented items (pseudowords, illegal strings), meaning that the model can accurately make lexical decisions. N for items in the model is computed only on the basis of the model's vocabulary. Units of mean semantic activation are arbitrary. Though there is no formal relationship between model time and real time, the first tick of model time does correspond to stimulus onset, and the end of the model's processing epoch corresponds roughly to the end of the N400 and onset of the LPC in the ERPs.

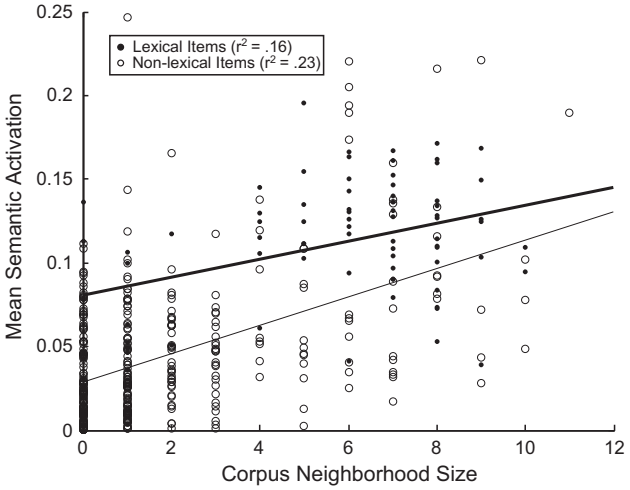

Fig. 7. Orthographic neighborhood size effect in single item model mean semantic activations. Lexical items (words and acronyms) are represented by filled dots, non-lexical items (pseudowords and illegal strings) are represented by empty dots. Note that the slopes representing the relationship between orthographic neighborhood size and mean semantic activation in the model are quite similar, though the intercepts differ. N for items in the model is computed only on the basis of the model's vocabulary. Units of mean semantic activation are arbitrary.

is a large main effect of N on mean semantic activation in the model ($\beta = .0085$, 95% confidence interval $.0069 < \beta < .101$), and no interaction between N and lexicality ($\beta = -.0031$, 95% confidence interval $-.0065 < \beta < .0002$). Unlike the ERP data, however, in the model there was a reliable main effect of lexicality ($\beta = .05$, 95% confidence interval $.0332 < \beta < .0712$). This is a direct result of aiming to produce a model capable of performing lexical decision—as, of course, if words and pseudowords (or acronyms and illegal strings) elicited identical mean amounts of semantic activation they would be impossible to tell apart on that signal.

We followed up the multiple regression with a focused analysis of N effects in the model, as these effects are particularly prominent in the ERPs. In the model, the single regression of N on mean semantic activation is strongly reliable for both represented items (words and acronyms: $r = .40$, $r^2 = .16$, $p < .0001$) and non-represented items (pseudowords and illegal strings: $r = .48$, $r^2 = .23$, $p < .0001$). If the regression is computed over all items (i.e., collapsed over lexicality), the amount of variance explained is comparable to the 30.6% of variance uniquely explained by N in the ERPs ($r = .61$, $r^2 = .37$, $p < .0001$. Fig. 7 presents the model regression data comparable to the ERP regression data presented in Fig. 4. Note that, just as in the ERP data, the slopes of the trendlines representing the relationship between N and mean semantic activation the model are very similar for represented vs. non-represented items (.005 vs. .008, respectively), though the intercepts are different, representing the model's ability to perform lexical decision.

3.3. Discussion

Simulation 1 served several goals. First, it helped to determine whether a PDP reading model with neurally plausible architecture could produce dynamics on its semantic output layer that resembled the N400 ERP component. In this the model was successful: the time course of mean semantic activation for words, pseudowords, acronyms, and illegal strings in the trained model strongly resembled N400 morphology in several critical ways. Namely, semantic output was delayed slightly from the onset of stimulus presentation (i.e., from the time when input was clamped on in the model)—just as the N400 does not onset immediately when a stimulus is presented. When activation began to arise in semantics, it did so in a way consistent with N400 morphology, by monotonically rising and falling into a stable state which was predictive of lexicality. This characteristic in the model is, in fact, not only consistent with N400 morphology but also with the morphology of subsequent components: the Late Positive Complex (LPC), which follows the N400, often displays a relatively tonic level of activation which has been shown to be predictive of the lexicality of the item which elicited it (e.g., Laszlo & Federmeier, 2009; Laszlo et al., in press).

Since the model was successful in producing N400-like dynamics in its output, the second goal was to explore the degree to which its simulated N400 activity resembled N400 activity in the single-item ERP corpus. Here, there were both similarities and differences between the model and the physiological data. Both the model and the physiological data displayed a strong effect of orthographic neighborhood size, with N in fact explaining similar amounts of variance in model and ERPs, and with no interaction between N and lexicality. Additionally, in both the model and the ERPs, the slope of the regressions of N on mean semantic activation (in the case of the model) or N400 mean amplitude (in the case of the ERPs) were highly similar for lexical and non-lexical items. However, one difference between model and ERPs emerged in these analyses: there was a main effect of lexicality in the model, with words and acronyms eliciting more semantic activity than pseudowords and illegal strings. This was not the case in the ERPs. Based on comparison with previous simulations, it is clear that this difference is largely the result of the model's ability to perform lexical decision solely on the basis of semantic output—nearly identical models which do not perform the LDT show exactly the same pattern as the ERPs (Laszlo & Plaut, 2011). The fact that it was able to make accurate lexical decisions on the basis of a simple fixed activation threshold—even without being explicitly trained on lexical decision—is important, as it demonstrates that the ERP model, which makes use of several neurally plausible architectural features and was primarily designed to simulate ERP data, is not completely divorced from the vast cognitive modeling literature on reading. Further, it suggests that a neurally plausible model is also a cognitively plausible one.

In sum, the model was largely successful in simulating the phenomena it aimed to simulate, and at demonstrating that a PDP model can simulate not only general properties of ERPs, but also specific, key results pertaining to the obligatory semantics view of N400 processing. In the model, even meaningless, illegal consonant strings elicited activation in semantics, graded by the similarity of those strings to represented items in the training corpus. Additionally, items regression analysis indicated that the relationship between N and mean semantic activation was quite similar for semantically represented items and items with out semantics. Each of these phenomena, when observed in the ERPs, have been interpreted as being consistent with PDP models, and the present simulations indicate that such an interpretation is warranted.

In light of the model's successes in Simulation 1, a clear question for additional exploration is: To what degree did the neurally plausible architecture of the ERP model contribute to its success? We investigate this question in Simulation 2, in which the neurally plausible features of the ERP model are removed. Specifically, the constraints on the separation of excitation and inhibition are removed: units in the second set of simulations have no constraints on the sign of their outgoing weights, or on the distribution of inhibitory connections. In what follows, we will refer to this model as the *unconstrained* model, while the version with excitation and inhibition separated will be referred to as the *constrained* model. The critical issue to be determined by the unconstrained model is this: will the model still display activation dynamics in semantics that resemble N400 morphology without its neurally plausible features?

4. Simulation 2

4.1. Methods

The unconstrained model was identical to the constrained model with the following exceptions: In the unconstrained model, there were no constraints on the sign of a unit's outgoing connections, with one consequence being that negative weights were allowed between levels of representation. Thus, all units could have outgoing connections of any sign. The unconstrained model received exactly the same amount of training as the constrained model: 3000 epochs of training on words and acronyms for the autoencoder, followed by 9000 epochs of training on words, acronyms, and wordlike nonwords for the semantic output layer.

4.2. Results

4.2.1. Autoencoder

After 3000 epochs of training, the unconstrained autoencoder's performance was perfect (100%).

4.2.2. Semantics

After 9000 epochs of training, the unconstrained network was 90% (398/441) accurate in producing correct semantics (in the case of words and acronyms), or in staying silent (in the case of pseudowords and illegal strings). Four errors were made for words, zero for acronyms, 19 for pseudowords, and 20 for illegal strings. Fig. 8 displays the item-aggregated mean activation in semantics for words, acronyms, pseudowords, and illegal strings. It is clear from Fig. 8 that semantic activation in the unconstrained network does not resemble N400 activation in at least one important respect: In the unconstrained network, there are two distinct peaks in semantics, one occurring quite early on in processing. The first peak is absent in both the constrained network and the ERPs.

Like the constrained model, the unconstrained model is able to perform accurate lexical decisions based on a set activation threshold. Using the same threshold as was adopted for the constrained model, the unconstrained model is 87% accurate in discriminating lexical from non-lexical items (384/441). Also like the constrained model, a simultaneous multiple regression on mean semantic activation with predictors of N, lexicality and the N x lexicality interaction revealed a main effect of N ($\beta = .0041$, 95% confidence interval $.0031 < \beta < .0050$), a main effect of lexicality ($\beta = .0425$, 95% confidence interval $.0312 < \beta < .0538$), and no interaction between the two ($\beta = -.0015$, 95% confidence interval $-.0035 < \beta < .0005$). Focused single regression analysis reveals that, as in the constrained model, there is a reliable correlation between N and mean semantic activation for both lexical items ($r = .31$, $r^2 = .10$, $p = .0055$) and non-lexical items ($r = .41$, $r^2 = .16$, $p < .0001$). Trendline slopes for lexical (.003) and non-lexical (.004) items are quite similar. If the regression is computed over all items (i.e., collapsed over lexicality), the amount of variance explained is comparable to the 30.6% of variance uniquely explained by N in the ERPs ($r = .58$, $r^2 = .34$, $p < .0001$).

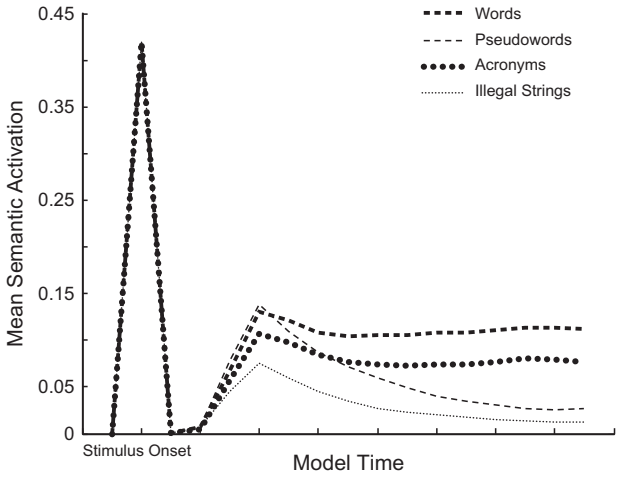

Fig. 8. Item-aggregated time course of semantic activation for words, pseudowords, acronyms, and illegal strings in the unconstrained model. The dynamics of activation the unconstrained model are much less similar to N400 morphology than those of the constrained model. Though there is no formal relationship between model time and real time, the first tick of model time does correspond to stimulus onset, and the end of the model's processing epoch corresponds roughly to the end of the N400 and onset of the LPC in the ERPs.

4.3. Discussion

On the whole, the unconstrained model performs very similarly to the constrained model, which is to be expected as they are nearly identical. What is especially important to note is that the *functional* effects present in the unconstrained model are the same as those in the unconstrained model: that is, both models display the ability to make lexical decisions, as well as reliable effects of orthographic neighborhood size in the absence of an interaction between N and lexicality. Where the results of the simulations differ is in the *dynamics* of semantic activation—semantic activation in the constrained model strongly resembles the morphology of the N400, while semantic activation in the unconstrained model does not. This is a clear example of the benefits of linking a time course of processing from ERPs with model dynamics: the ERPs rule out a model which exhibits all the appropriate functional effects but which does not display the correct internal dynamics. Constraint on internal dynamics of a model such as that which rules out the unconstrained simulation would not be available from behavioral data alone. The unique contribution of the ERPs as target phenomena is that they strongly constraint the internal dynamics of potential models to a degree not possible from end state behavioral data.

One issue left unresolved by Simulation 2 is whether it was a lack of feedforward inhibition, lack of within-level mixed connections (i.e., units with both positive and negative outgoing weights), or both of these that are required to produce the dynamics in semantics observed in the ERPs and in Simulation 1. We investigated this matter with two follow-up simulations: the *between* simulation, which was identical to Simulation 1 but allowed inhibition between levels of representation, and the *within* simulation, which allowed mixed-sign connections within a level of representation. Neither of these simulations produced the appropriate dynamics in semantics: in fact, both were quite poor at discriminating between item types (i.e., the effect of N was absent in the between simulation, and extremely small in the within simulation).[1] Thus, neither limiting inhibition to within a level of representation or restricting units in the sign of their outgoing connections alone is enough to produce the dynamics observed in Simulation—both together are required.

5. General discussion

The present simulations represented an exploration into the feasibility of simulating ERP reading data in a PDP model. We examined the specific case of visual word recognition data from the single-item ERP corpus—data that has explicitly been cast as conforming to PDP principles. Because ERPs represent summed excitatory and inhibitory post-synaptic potentials, we implemented basic principles of cortical excitation and inhibition in the architecture of the model. Specifically, in our constrained model, we did not allow individual units to have both excitatory and inhibitory outgoing connections, we limited the number of inhibitory units, and we only allowed relatively short-range inhibition. The constrained model was able to successfully simulate a number of key effects from the single-item ERP corpus, most importantly very similar effects of orthographic neighborhood size on items which had been trained (words and acronyms) and items which had not been trained (pseudowords and illegal strings), in the absence of an interaction between N and lexicality. The correspondence between these features of the simulations and the same effects in the ERPs represent converging evidence for the obligatory semantics view of N400 processing. In addition, the model was able to make accurate lexical decisions based on a set activation threshold, thus extending and broadening the scope of preliminary modeling work focused on the ERP data only (Laszlo & Plaut, 2011). The constrained model simulated all these phenomena in the context of internal dynamics which resembled the morphology of the N400 ERP component. As discussed above, when excitation and inhibition were no longer handled in a neurally plausible fashion, the model was still able to produce many of the key functional results (e.g., the effect of N, accurate lexical decisions), but no longer displayed dynamics in semantics that were consistent with the ERPs. Thus, the veridical manner in which excitation and inhibition were handled in the constrained model were at least partly responsible for producing N400-like dynamics—a result that is encouraging although perhaps not surprising given the neural source of the N400 signal. The fact that a functionally adequate model was ruled out on the basis of striking differences between its internal dynamics and those present in the ERPs is representative of the unique contribution that ERP data can make as target data for simulations of visual word recognition.

The internal dynamics of the constrained model were consistent with the obligatory-semantics view of N400 processing in critical respects. As predicted by the obligatory-semantics view, nonwords in the model made clear contact with semantics. This was true even for illegal consonant strings, which were never trained to be linked with semantics and which had low neighborhood sizes within the model's vocabulary. This observation in the model is consistent with the interpretation that N400 effects observed for illegal strings (e.g., Laszlo & Federmeier, 2011, Laszlo et al., in press) represent the obligatory contact that illegal strings make with semantics. In contrast, this behavior in the model is inconsistent with theories of the N400 that suggest it responds selectively to items with regular spelling-sound correspondences (e.g., Deacon, Dynowska, Ritter, & Grose-Fifver, 2004), or theories which suggest that N400 processing takes place only after an input has been uniquely identified as a particular lexical item (e.g., Hagoort, Baggio, & Willems, 2009). In fact, insofar as the model constitutes evidence that contact with semantics is made prior to unique lexical selection, it is more generally inconsistent with models outside of the ERP literature which implement a strong "lexical stage", such as the Entry Opening Model (see Forster & Hector, 2002). Instead, the evident contact with semantics made even by illegal strings in the model constitutes converging evidence for both the obligatory semantics view in particular and cascaded models of visual word recognition more generally (e.g., Plaut & Booth, 2000).

One important difference between the behavior of the constrained model and the ERPs was the mean difference in activation between lexical items and non-lexical items. This difference is a consequence of the model's ability to perform lexical decisions accurately solely on the basis of mean semantic activation—it does not exist in similar models which do not perform lexical decision (Laszlo & Plaut, 2011). Though insisting that the model to perform lexical decision as a criterion for success caused it to deviate from the observed ERP dynamics—especially at the end of the processing epoch—the model's ability to perform lexical decision while still simulating the several key ERP effects constitutes an improvement over previous work (Laszlo & Plaut, 2011). Because the model was ultimately given the task of performing lexical decision and participants in the single-item study were not—instead monitoring the stimulus stream for proper names, one obvious question is whether a lexicality effect might emerge on the N400 if participants were performing lexical decision. In fact, we have conducted such a study, where participants were presented with the exact same items used in the single-item study, but were asked to make modified lexical decisions about them (responding that names, words, and acronyms were "familiar" and pseudowords and consonant strings were not; Laszlo, Stites, & Federmeier, 2010). Interestingly, when using the standard mean amplitude over a broad time window analysis technique typically employed in ERP studies, no effect of lexicality was observed on the N400 even in that study, again appearing only on the LPC (Laszlo et al., 2010, in press). In the model, the difference in mean semantic activation between lexical items and non-lexical items is most pronounced at the end of the processing epoch, when the model has come to a relatively stable level of activation (see Fig. 6). One potential interpretation of the modeling data is thus that the processing occurring during the N400 terminates with a stable representation of the semantics of an item (or its lack of semantics), and this representation is fed forward to LPC processing as a basis on which, for example, lexical decisions can be made (see Laszlo et al., in press). Because the stable difference between lexical and non-lexical items occurs primarily at the end of N400 processing, and because LPC processing follows the N400 directly in time and tends to have a quite similar scalp distribution, the stable differences between lexical and non-lexical items that the model suggests are part of terminal N400 processing could easily be missed or obscured by the LPC—especially given the fact that N400 effects are typically analyzed via component mean amplitude, a measure that could easily miss small effects that occur only in a limited portion of

[1] The details of these follow-up simulations are available from the first author on request.

the epoch on which the mean is taken. It is particularly difficult to disentangle N400 and LPC effects because of their temporal contiguity and similar scalp distributions—what would be needed to investigate the hypothesis about the N400 outputting a stable semantic signal to the LPC would be high density ERP recordings of responses to the same items investigated here, analyzed at fine temporal intervals.

That the model has suggested the existence of subcomponents of N400 processing that may previously have been obscured by typical ERP data analysis techniques is a good example of how models can help to move theories in the ERP literature forward, just as ERP results constrained the internal dynamics of the model in the current simulations. This sort of reciprocal relationship between modeling and cognitive neuroscience is an important reason to interweave modeling with cognitive neuroscience investigations even more tightly in the future. In service of this goal, it is important to note that the neurally plausible architecture employed here is not specific to reading models—it could naturally be employed in cognitive models of essentially any phenomenon. All that is required is setting appropriate constraints on the sign of outgoing weights in any model, and limiting the number of available inhibitory units.

The model constitutes a step forward in the goal of linking ERP data, neuroscience, and PDP modeling. However, there still remains a substantial amount of work to be done both in increasing the neural plausibility of the model, simulating more nuanced characteristics of the ERP data, and making more contact with the behavioral literature. In terms of neural plausibility, the model makes use of back-propagation to reduce error during training and effect learning. However, back-propagation is considered unlikely as a mechanism of neural learning (e.g., O'Reilly, 1996b). Thus, in moving forward, it will be advantageous to investigate the degree to which the ERP model can produce appropriate results if trained with a more biologically plausible learning algorithm, such as Contrastive Hebbian Learning (Ackley, Hinton, & Sejnowski, 1985), which at least in some cases provides similar solutions to back-propagation (e.g., Xie & Seung, 2003), while avoiding many of back-propagation's biologically implausible properties. Intermediate algorithms, which combine the power of back-propagation with the neuronal validity of Hebbian learning (e.g., O'Reilly, 1996a), may prove useful in bridging the gap between the two techniques.

Another possibility for immediate improvement of the model's neural plausibility is found in the way fast and slow inhibition is approximated in the model. Currently, a single inhibitory unit provides both fast and slow inhibition, through use of the multi-linear inhibition function. However, in the cortex, single inhibitory neurons do not vary drastically in their time constants. Instead, separate populations of neurons provide fast and slow inhibition (Benado, 1994; Traub et al., 1989). In future versions of the model, it would be quite simple to have separate inhibitory units with separate time constants—for example one fast inhibitory unit and one slow inhibitory unit. This can easily be accomplished by implementing different time constants of integration on different inhibitory units, or by implementing different slopes on the activation functions of different inhibitory units.

In terms of simulating more nuanced aspects of the ERP data, the model's training corpus in the present simulations included words which essentially varied only in terms of orthographic neighborhood size: they were all the same length, all the same frequency (as each other, though they were more frequent than the wordlike nonwords they were trained with), and though they differed in neighbor frequency, neighbor frequency was equivalent to N in the corpus (since all items were the same frequency). Similarly, since semantic features were assigned to each item randomly, there was no coherence to the semantic structure of the corpus, meaning that there were no meaningful correlates in the model to variables such as strength of lexical association. However, all of these variables have prominent effects in the single-item ERP corpus (Laszlo & Federmeier, 2011)—we focused on N here only because of the strength and robustness of its effects. In moving forward, it will be important to develop a training corpus with lexical characteristics more similar to those of the items in the single-item ERP corpus. The development of more realistic semantic representations, in particular, will be critical if the model is to be extended beyond simulation of items in unconnected lists, to simulation of ERP effects observable in sentence comprehension. The N400 is known to be extremely sensitive to even very fine manipulations of factors such as sentence constraint (e.g., Federmeier, Wlotko, De Ochoa-Dewald, & Kutas, 2007), making sentence-level N400 effects an important venue for future modeling work. In addition to more veridical semantic features, development of a sentence-level ERP model would require a mechanism for accruing semantic context over time. A clear starting place for such a mechanism would be to simply not reset semantic activations to zero between each stimulus presentation, as was done in the current simulations.

Finally, just as it will be important to develop more detailed simulations of the abundance of effects observable in the single-item ERPs, it will also be important to improve the sophistication of the model's cognitive simulations. Presently, it is able to make accurate lexical decisions, but there are many more benchmark behavioral phenomena—some of them even pertaining to more detailed aspects of lexical decision—that the model has not attempted to address—for example, the frequency and consistency effects in lexical decision, the word superiority effect, and semantic categorization effects. Increasing the model's behavioral sophistication will be critical to bringing it into better contact with past models of reading—a contact which is desirable because of the substantial insights into representation and processing already available from the model's thematic predecessors (e.g., Harm & Seidenberg, 2004; Plaut et al., 1996; Seidenberg & McClelland, 1989). For example, the model's predecessors also consider the interaction of phonological representations with semantics and orthography—with phonology's role being extremely important in interpretation of many behavioral findings (e.g., pseudohomophone effects) as well as patterns of impairment in dyslexia. The ERP model currently has no implemented phonology, but adding appropriate phonological representations is likely to be important for expanding the model's scope.

Acknowledgments

The authors acknowledge K.D. Federmeier and R.C. O'Reilly for their insightful discussion of the unique challenges involved in simulating the ERP data. This Research was supported by NIMH T32 MH019983 to Carnegie Mellon University and NICHD F32 HD062043 to S.L. Reprint requests should be directed to S.L.: cogneuro@alum.mit.edu.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science, 9*, 147–169.

Barber, H. A., & Kutas, M. (2007). Interplay between computational models and cognitive electrophysiology in visual word recognition. *Brain Research Reviews, 53*, 98–123.

Benado, L. S. (1994). Separate activation of fast and slow inhibitory postsynaptic potentials in rat neocortex in vitro. *The Journal of Physiology, 476*, 203–215.

Coltheart, M., Davelaar, E., Jonasson, J., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance IV* (pp. 535–555). Hillsdale, NJ: Erlbaum.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological review, 108*, 204–256.

Crick, F., & Asanuma, C. (1986). Certain aspects of the anatomy and physiology of the cerebral cortex. In D. E. Rumelhart, J. L. McClelland, & PDP Research Group (Eds.). *Parallel distributed processing: Explorations in the microstructure of cognition: Psychological and biological models* (Vol. 2). Cambridge: MIT Press.

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

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

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

Forster, K. I., & Hector, J. (2002). Cascaded versus noncascaded models of lexical and semantic processing: The *turple* effect. *Memory and Cognition, 7*, 1106–1117.

Grainger, J., & Holcomb, P. J. (2009). Watching the word go by: On the time-course of component processions in visual word recognition. *Language and Linguistics Compass, 3*, 128–156.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences, 14*, 357–364.

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

Hagoort, P., Brown, C. M., & Swaab, T. Y. (1996). Lexical-semantic event-related potential effects in patients with left hemisphere lesions and aphasia, and patients with right hemisphere lesions without aphasia. *Brain, 119*, 627–649.

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

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review, 111*, 662–720.

Hinton, G. E. (1989). Connectionist learning procedures. *Artificial Intelligence, 40*, 185–234.

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

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

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

Kutas, M., & Hillyard, S. A. (1983). Event-related brain potentials to grammatical errors and semantic anomalies. *Memory and Cognition, 11*, 539–550.

Kutas, M., & Hillyard, S. A. (1984). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207*, 203–205.

Laszlo, S., & Plaut, D. C. (2011). Simulating event-related potential reading data in a neurally plausible parallel distributed processing model. In *Proceedings of the 33rd annual conference of the cognitive science society*, Lawrence Erlbaum Associates, Mahwah, NJ.

Laszlo, S., Stites, M., & Federmeier, K. D. (in press). Won't get fooled again: An event-related potential study of task and repetition effects on the semantic processing of items without semantics. *Language and Cognitive Processes*.

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

Laszlo, S., & Federmeier, K. D. (2008). Minding the PS, queues, and PXQs: Uniformity of semantic processing across multiple stimulus types. *Psychophysiology, 45*, 458–466.

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

Laszlo, S., & Federmeier, K. D. (2011). The N400 as a snapshot of interactive processing: Evidence from regression analyses of orthographic neighbor and lexical associate effects. *Psychophysiology, 48*, 176–186.

Laszlo, St., Stites, M., & Federmeier, K. D. (2010). Task and repetition effects on the semantic processing of items without semantics. *Psychophysiology, 47*(1), S28.

McCarthy, G., Nobre, A. C., Bentin, S., & Spencer, D. D. (1995). Language-related field potentials in the anterior-medial temporal lobe: I. Intracranial distribution and neural generators. *Journal of Neuroscience, 15*, 1080–1089.

Nobre, A. C., & McCarthy, G. (1995). Language-related field potentials in the anterior-medial temporal lobe: II. Effects of word type and semantic priming. *Journal of Neuroscience, 15*, 1990–1998.

O'Reilly, R. C. (1996a). *The Leabra model of neural interactions and learning in the neocortex*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA.

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

Perry, C., Ziegler, J. C., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: The CDP+ model of reading aloud. *Psychological Review, 114*, 273–315.

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

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*, 56–115.

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

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

Traub, R. D., Miles, R., & Wong, R. K. (1989). Model of the origin of rhythmic population oscillations in the hippocampal slice. *Science, 243*, 1319–1325.

Tse, C.-Y., Lee, C.-L., Sullivan, J., Garnsey, S. M., Dell, G. S., Fabiani, M., et al. (2007). Imaging cortical dynamics of language processing with the event-related optical signal. *Proceedings of the National Academy of Sciences of the United States of America, 104*, 17157–17161.

van Berkum, J. J. A. (2008). Understanding sentences in context: What brain waves can tell us. *Current Directions in Psychological Science, 17*, 376–380.

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

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

White, E. L. (1989). *Cortical circuits: Synaptic organization of the cerebral cortex, structure, function, and theory*. Boston: Birkhauser.

Xie, X., & Seung, H. S. (2003). Equivalence of backpropagation and contrastive Hebbian learning in a layered network. *Neural Computation, 15*, 441–454.