Simulating Event-Related Potential Reading Data in a Neurally Plausible Parallel Distributed Processing Model

Sarah Laszlo (sarahlaszlo@cmu.edu)
Department of Psychology and Center for the Neural Basis of Cognition, 5000 Forbes Ave
Pittsburgh, PA 15213 USA

David C. Plaut (plaut@cmu.edu)
Department of Psychology and Center for the Neural Basis of Cognition, 5000 Forbes Ave
Pittsburgh, PA 15213 USA

Abstract

Parallel Distributed Processing (PDP) models have always been considered a particularly likely framework for achieving neural-like simulations of cognitive function. To date, however, minimal contact has been made between PDP models and physiological data from the brain performing cognitive tasks. We present an implemented PDP model of Event-Related Potential (ERP) data on visual word recognition. Simulations demonstrate that a novel architecture with improved neural plausibility is critical for successfully reproducing key findings in the ERP data.

Keywords: Parallel Distributed Processing (PDP), Event-Related Potentials (ERPs), visual word recognition

Introduction

From their initial development, PDP models have been considered an especially promising framework for building simulations which perform cognitive tasks with a mechanism similar to that employed in the brain (e.g., McClelland, Rumelhart, & Hinton, 1986). This optimism derives in large part from the fact that the basic processing units in PDP models are neuron-like, in that the models typically employ many interconnected units, each performing relatively simple computations, and represent information in a distributed fashion (c.f., Bowers, 2009; Plaut & McClelland, 2010).

The sense that PDP models should lend themselves well to simulating data from cognitive neuroscience—that is, brain data relating to cognitive function—is not only historical. Indeed, especially in the domain of single word reading, it is currently common for descriptions of prominent models to suggest that improvements over existing models could and should be made by increased contact with data from cognitive neuroscience (e.g., Harm & Seidenberg, 2004; Perry, Ziegler, & Zorzi, 2007). Correspondingly, as theories of how reading works based on neuroimaging data have become increasingly well-specified, a consensus is emerging—especially in the Event-Related Potential (ERP) literature—that interpretation of brain data could benefit from the guidance of formal computational models (e.g., Banquet & Grossberg, 1987; Barber & Kutas, 2007; Van Berkum, 2008). For example, one currently viable theory of the functional significance of the N400 ERP component (a centro-posterior component peaking around 400 ms post stimulus onset, and thought to reflect lexical-semantic access: see Kutas & Federmeier, in press, for review) suggests that N400 activity represents the continuous activation of semantic features associated with an orthographic input at either a whole or partial item level (e.g., the activation of the semantic features associated with both FORK and PORK in response to presentation of the word FORK; Laszlo & Federmeier, 2011). Under this so called *obligatory semantics* view, contact with semantics is made automatically by every orthographic input, and interaction between levels of representation is continuous (explaining, for example, sentence context effects on illegal nonwords; Laszlo & Federmeier, 2009).

Two features of this theory are particularly relevant for implementation in a computational model. First, the proposal that orthographic sub-parts of items can activate the semantic features of orthographically similar items extends to nonwords, such that pseudowords (e.g., GORK) and even consonant strings (e.g., XFQ) are allowed to contact semantics —explaining robust N400 effects observed for these items (e.g., Laszlo & Federmeier, 2009). This feature of the obligatory semantics view implicates a word recognition system that is not strongly lexicalized, and as such would seem to be more appropriate for simulation in a distributed PDP framework than in competing frameworks with explicit lexical representations (cf., Perry, Ziegler, & Zorzi, 2007). Though it would be possible for lexicalized models to account for these data by simply allowing very un-wordlike nonwords to activate their neighbors weakly, such a system is no longer strongly lexicalized in that its internal response to each input involves the activation of a number of units, with that activation graded by similarity to the input-- a system that is essentially distributed. Second, the continuous, interactive nature of the obligatory semantics view stronly contrasts with staged models of word recognition (e.g., Borowsky & Besner, 1993). Thus, the obligatory semantics view posits a mechanistic account of visual word recognition resonant with the PDP approach, but the question remains: would an implemented PDP model exhibit the patterns of effects in the ERP data suggestive of a non-lexicalized, continuous system (e.g., N400 effects for illegal consonant strings)?

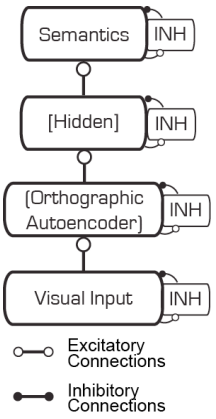

Figure 1: Architecture of the ERP model. Lines with empty circles indicate excitatory connections, lines with filled circles indicate inhibitory connections. INH stands for "inhibitory". Note that no units have both excitatory and inhibitory outgoing connections, and that inhibition is always within, never between, levels of representation.

It seems clear, then, that converging evidence for the obligatory semantics view could be provided by simulations of a model instantiating its key theoretical constructs. In addition to the added support for a particular view of N400 processing that could be provided by a successful simulation, the effort to model the continuous, internal dynamics of reading as reflected in the psychophysiology could be beneficial for improving reading models as well. The types of measures with which reading models are typically concerned—for example, reaction time (RT) and accuracy—provide important information about the reading system. However, RT and accuracy are fundamentally end-state measures, and therefore do not provide *direct* evidence about the continuous, internal processing involved in reading between when an item is presented and when a response is made. ERPs provide information about exactly such processing, and thus have the potential to provide important constraints on the internal functioning of reading models.

There are both empirical and computational challenges to producing a large-scale reading model that can simulate the ERP data. In the empirical domain, although it would be advantageous to simulate item-level effects, until very recently no ERP data suitable for modeling were available, as the low numbers of participants typically run in ERP reading studies prohibited the formation of stable item ERPs. However, the recent advent of a massive corpus of single-item ERP data (Laszlo & Federmeier, 2011) collected specifically for the purpose of informing a computational model has effectively addressed this issue.

In particular, one largely theoretical and one implementational challenge are fundamental to an attempt to build a model of the ERP reading data. At the theoretical level, it is necessary to determine what parameter of the model should be linked with the dependent measure in the ERP data: amplitude of the N400 component. Implementation-wise, because ERPs fundamentally reflect synchronous excitatory and inhibitory post-synaptic potentials in the cortex, it is especially important to handle excitation and inhibition in the network in a manner true to the way they are handled in the brain. We next discuss our approach to each of these challenges, before summarizing the critical ERP data, describing the ERP model, and presenting simulations.

A Linking Hypothesis: N400 to Model

Past computational models of reading have been solely concerned with simulating behavioral RT or accuracy data (e.g., Harm & Seidenberg, 2004; Perry, Ziegler, & Zorzi, 2007; though see also Banquet & Grossberg, 1987 for joint ERP and computational work in another domain). Thus, parameters of past models selected for comparison with the empirical data have been linked to RT or accuracy—for example, number of processing cycles to settling in the case of RT. However, RT or accuracy measures are not appropriate for testing theories about the ERP data, as the critical measure in the relevant ERP studies is N400 mean amplitude, often in the absence of any explicit response to the eliciting stimulus.

The determination of what parameter of the model to link with N400 mean amplitude was guided largely by a pervasive pattern of effects in the single item ERP corpus suggesting that N400 amplitudes are larger for items which might reasonably be thought to elicit more overall activity in semantics. That is, items with larger orthographic neighborhood sizes, higher frequency of orthographic neighbors, more lexical associates, and higher frequency lexical associates all elicited larger N400s than did items with lower values on these measures (Laszlo & Federmeier, 2011). In addition to being pervasive in the single-item ERP corpus, this pattern of results is generally consistent with past work in factorial ERP designs (e.g., Holcomb, Grainger & O'Rourke, 2002; Laszlo & Federmeier, 2007), all of which point to N400 mean amplitude as a rough indicator of amount of semantic activation elicited by a target item-- at least when items are presented in random lists. Thus, we chose to link mean amount of semantic activation across the entire time course of processing in our network with N400 mean amplitude.

Excitation and Inhibition

ERPs reflect the synchronous firing of excitatory and inhibitory post-synaptic potentials in open-field configurations in the cortex. Thus, the neuroanatomy of excitation and inhibition is highly relevant to the final morphology of ERPs measured at the scalp. We considered two critical characteristics of excitation and inhibition in the cortex when planning the architecture of the ERP model. First, neurons in the cortex are either excitatory or inhibitory, but not both (e.g., Crick & Asanuma, 1986). Second, inhibitory connections are relatively short-range, with connections between cortical areas being largely excitatory (Crick & Asanuma, 1986). Neither of

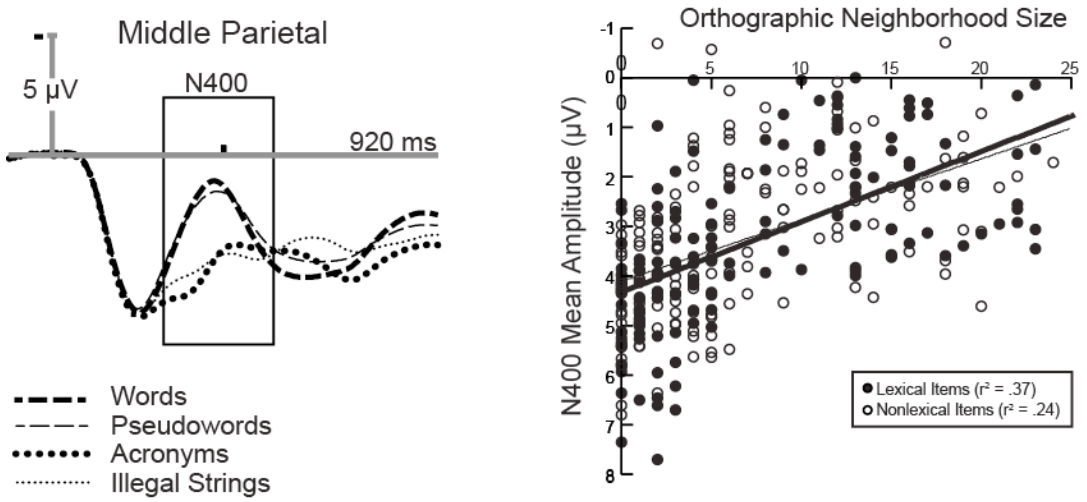

Figure 2: Left, item aggregated ERPs from the middle parietal electrode site representing the response to words, pseudowords, acronyms, and illegal strings. The N400 window is boxed. Right, scatter plot depicting the relationship between N400 mean amplitude and orthographic neighborhood size for all 300 single item ERPs. Lexical items (words and acronyms) are in filled circles, nonlexical items (pseudowords and illegal strings) are in empty circles. In both panels, negative is plotted up.

these characteristics have been implemented in past PDP reading models, as the typical architecture of such models allows for individual units to have both excitatory and inhibitory outgoing connections, and for inhibitory connections to exist between levels of representation (e.g., Plaut, McClelland, Seidenberg, & Patterson, 1997; Harm & Seidenberg, 2004). In the ERP model, we sought to move towards a more neurally plausible architecture by separating excitation and inhibition, and by only allowing excitatory connections between levels of representation (a strategy which has also been suggested for cognitive models; see for example Grossberg, 1984).

As can be observed in Figure 1, which displays the architecture of the ERP model, this was accomplished by pairing each bank of excitatory-only units within a level of representation with a bank of inhibitory-only units within that same layer. Thus, no unit had both excitatory and inhibitory projections: only one or the other. Only the excitatory units were allowed projections between layers of representation, meaning that the range of inhibitory connections was limited to within a level of representation. Each inhibitory layer included far fewer units than its matching excitatory layer (there were only 6.6% as many inhibitory units as excitatory units, over all), in accordance with the fact that the large majority of neurons in the cortex are excitatory (e.g., White, 1989). As we shall see, separating excitation and inhibition in this fashion is critical for successfully simulating the ERP data.

Method: ERPs

ERPs were acquired from 120 participants (58 female, age range 18-24, mean age 19.1) who viewed an unconnected stream of text consisting of words (e.g., HAT, MAP), acronyms (e.g., VCR, AAA), pseudowords (e.g., TUL, KAK), illegal strings (e.g., CKL, KKB), and names (e.g., SARA, DAVE). Words, acronyms, pseudowords, and illegal strings were used as the single-item study was a replication of a previous study using these same item types (Laszlo & Federmeier, 2007) and we wanted to be certain before collecting 120 participants worth of data that our items and task were already well studied. Names were of no experimental interest but served as the putative targets in the experiment: Participants were required to press a button each time a name appeared (this was the case in Laszlo & Federmeier, 2007, as well). Single item ERPs were formed by averaging, across participants but not items, at each electrode time-locked to the onset of each word, pseudoword, acronym, or illegal string. Three-hundred single item ERPs were thus formed at each electrode, 75 in each of the 4 critical item types. In addition, more traditional ERPs representing the averaged within-subject response to, for example, all words, were also computed. For a more detailed description of the ERP methods, see Laszlo & Federmeier, 2011.

Results: ERPs

Automated large scale multiple regression analyses conducted on N400 mean amplitudes for all 300 single item ERPs revealed that when all items were included (i.e., not just lexically represented items such as word and acronyms) by far the largest predictor of N400 amplitude was orthographic neighborhood size (Coltheart's N, the number of words that can be produced by changing 1 letter of a target item): N

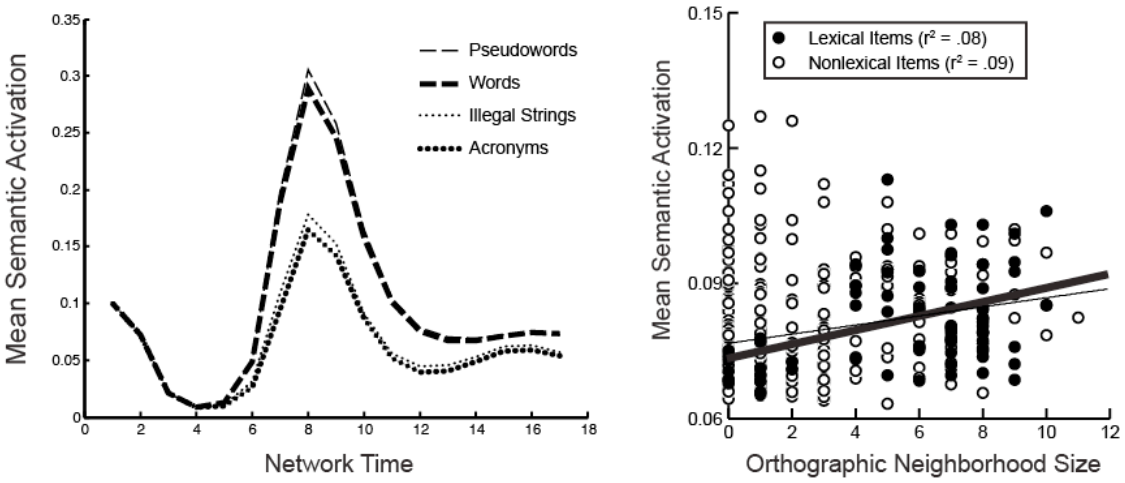

Figure 3: Left, mean semantic activation in the constrained network over time, for pseudowords, words, illegal strings, and acronyms. Note that high N items (words, pseudowords) tend to elicit more activation than low N items (acronyms, illegal strings), regardless of lexicality. Right, scatter plot depicting the relationship between mean semantic activation and orthographic neighborhood size for all 441 individual items in the model's testing set. Lexical items (words and acronyms) are in filled circles, nonlexical items (pseudowords and illegal strings) are in empty circles.

explained 30.6 % of unique variance in N400 amplitude, followed by summed frequency of orthographic neighbors, which explained only 1.2 % of additional variance. A 2 x 2 items analysis of variance (ANOVA) with factors of N (high or low) and lexical type (lexical: words and acronyms, nonlexical: pseudowords and illegal strings) revealed a main effect of N ($F_{1,296} = 159.7$, $p < .0001$) but no effect of lexical type ($F = .19$), and no interaction between the two ($F = 1.1$) Figure 2 displays the N effect in the single item ERPs for both individual items and categories of items—the N effect manifests itself in the item aggregated ERPs as words and pseudowords (i.e., high N items) eliciting larger N400s than acronyms and illegal strings (i.e., low N items). In these data, as in past studies with similar items (e.g., Laszlo & Federmeier, 2007; 2009), individual lexical characteristics--such as N--are considerably stronger predictors of N400 amplitude than categorical labels such as lexical class (e.g., words v. acronyms).

The prominence of the N effect, combined with previous results demonstrating that, unlike effects of other lexical variables such as frequency or concreteness, N effects on the N400 are not eliminated by either repetition (Laszlo & Federmeier, 2007), or sentence context (Laszlo & Federmeier, 2009), suggests that N effects in the ERPs, in addition to being quite large, are also potentially of fundamental importance. For these reasons, and because of space considerations, we focus on N effects in the simulations presented below.

Method: Simulations

The architecture of the ERP model is depicted in Figure 1. A 15-unit visual input layer represents the visual features of each of three letters in 5 non-overlapping slots. The visual input layer feeds into an orthographic autoencoder, which was trained to reproduce the visual input. The autoencoder feeds through a 20-unit hidden layer to a 50-unit semantic layer, where relatively sparse representations (i.e., either 3 or 7 units) were trained to be associated with each visual input. Connections between level of representation are positive-only. Each level of representation (input, autoencoder, hidden layer, and semantics) has an associated inhibitory bank, connected as depicted in Figure 1. The logistic function is used to compute unit activations.

Training was accomplished by back-propagating cross-entropy error through time. The network was trained on 77 items (62 words and 15 acronyms). On each training trial, the visual input for one of the 77 items was clamped on, and activation was allowed to propagate through the network for 12 time steps with no accumulation of error. Targets were then presented for an additional 4 time steps. When training was complete, the network was tested on 441 items: the 62 words and 15 acronyms it was trained on, in addition to 279 illegal strings and 85 pseudowords which the network was not exposed to during training. The target for all illegal strings and pseudowords was for all semantic units to remain off.

Results: Simulations

At the end of training, the network was tested for its accuracy in producing the correct outputs in response to both the 77 inputs it was trained on (62 words, 15 acronyms), and the 364 additional nonwords that it was not exposed to in training (85 pseudowords, 279 illegal strings). An item was judged correct if the Euclidean distance between its actual output vector and its target vector was less than the distance between its actual output and any other target. Under this criterion, the network was 85% accurate (376 / 441 items correct). Errors largely

consisted of units being too weakly active for words and acronyms—this was a result of inhibition needing to be very strong in order to correctly turn all units off for pseudowords and illegal strings. Of critical importance was comparing the internal dynamics of the model to the ERP data. Figure 3 displays the simulation data corresponding to both the single item and item-aggregated ERP data. Regression analysis revealed that, just as in the ERPs, there was a strong relationship between N and mean semantic activation in the model (r = .34, p < .0001). Because there were not equal numbers of items in each lexical type cell in the simulations, it was not possible to perform ANOVA analyses corresponding to those performed on the ERP data. However, non-paramentric rank sum tests were able to confirm that, as it appears in Figure 3, words elicited more semantic activation than did acronyms or illegal strings (for acronyms, p < .0001; for illegal strings, p < .0001), but not more than did pseudowords (p = .82). Similarly, pseudowords elicited more semantic activation than did acronyms or illegal strings (for acronyms, p < .0001; for illegal strings, p < .0001). Acronyms and illegal strings also did not differ in the mean amount of semantic activity elicited (p = .80). Thus, the same pattern of effects was observed in the model as in the ERPs.

A second goal of the simulations was to determine whether the separation of excitation and inhibition was critical for producing the appropriate internal dynamics in the model. Therefore, a second simulation was run which was identical to the first but which did not place constraints on the sign of any connections (thus allowing units to have both excitatory and inhibitory connections simultaneously, and allowing between level of representation inhibition.) In what follows, we will refer to this as the *unconstrained* network, while the original network will be referred to as the *constrained* network. Figure 4 displays the item-aggregated results of this simulation.

After an identical amount of training, the unconstrained network was approximately as accurate as the constrained model, producing correct outputs for 83% of items (367 / 441). However, despite the similar level of overall performance, the internal dynamics of the unconstrained network did not resemble the critical ERPs. While there was still a relationship between N and mean amount of semantic activation in this simulation (r = .26, p < .0001), the pattern of this effect across lexical types did not match the empirical findings. For example, while words and pseudowords still elicited more activity than did acronyms (for words, p < .0001; for pseudowords, p < .0001), so did illegal strings (p = .035.), and it is clear that differences among item types, where they exist at all in this simulation, exist in relatively late tonic activation levels, as opposed to in the early sweep of over-activation observed in both the ERP data and the constrained simulation. Thus, a network which satisfactorily reproduces key dynamics of the ERP data when constrained to handle excitation and inhibition in a neurally plausible fashion does not do so when those constraints are removed.

Discussion

In our attempt to begin to bring computational formalism to a theory of visual word recognition from the ERP literature, we were able to successfully simulate critical findings from the single item ERP corpus in a PDP model which instantiated components of the obligatory semantics view of N400 processing. This success was at least in part due to the attention paid in the simulation to the neuroanatomy of excitation and inhibition, without which the successful model was not able to correctly reproduce the dynamics observed in the ERPs. The work consists of a proof of the concept that ERP data can successfully be simulated within the PDP framework, and provides a foundation for more comprehensive modeling of psychophysiological processes.

Though an encouraging initial attempt, it is clear that many refinements to the model are necessary avenues for future work. Two seem especially important: on the physiological side, extending the neural plausibility of the model, and on the behavioral side, making use of extensive past cognitive simulations to improve the model's contact with behavioral data.

While the architecture of the ERP model does represent an improvement in neural plausibility over past models, there are further improvements to be made. For example, as was described in the methods, the ERP model employs back-propagation through time to effect error reduction during training, despite the fact that back-propagation is considered unlikely as a mechanism of neural learning (e.g., O'Reilly, 1996). In response to this issue, current work with the ERP model focuses on successfully completing training with the

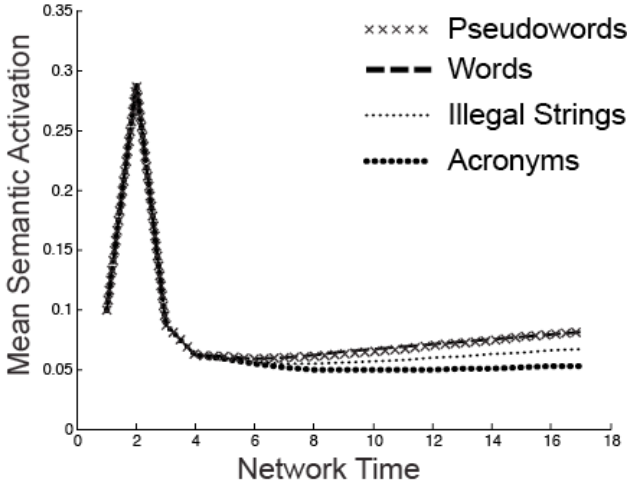

Figure 4: Mean semantic activation over time in the unconstrained network for pseudowords, words, illegal strings, and acronyms.

Contrastive Hebbian Learning algorithm (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.

The future development of the ERP model will also be guided by the strengths of past, related models of cognitive phenomena. The ERP model's direct predecessors, the so-called "triangle" models (e.g., Plaut, McClelland, Seidenberg, & Patterson, 1996; Harm & Seidenberg, 2004), have provided significant insight into the representations and flow of information in use in fundamental cognitive tasks such as lexical decision, semantic categorization, and the word superiority effect. Further development of the ERP model will make use of those insights, as in ongoing work we will consider the simultaneous simulation of ERP and behavioral data an important success criterion.

Acknowledgments

The authors acknowledge K.D. Federmeier and E.W. Wlotko for thoughtful discussion. This research was supported by NIMH T32 MH019983 to CMU and NICHD F32 HD062043 to S.L.

References

Ackley, D.H., Hinton, G.E., & Sejnowski, T.J. (1985). A Learning Algorithm for Boltzmann Machines. *Cognitive Science,* 9, 147-169.

Banquet, J.P., & Grossberg, S. (1987). Probing cognitive processes through the structure of event-related potentials during learning: An experimental and theoretical analysis. *Applied Optics,* 26, 4931-4946.

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

Borowsky, R., & Besner, D. (1993). Visual word recognition: A multistage activation model. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 19, 813-840.

Crick, F., & Asanuma, C. (1986). Certain Aspects of the Anatomy and Physiology of the Cerebral Cortex. In D.E. Rumelhart, J.L. McClelland, & The PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition Volume 2: Psychological and Biological Models.* Cambridge: MIT Press.

Grossberg, S. (1984). Unitization, automaticity, temporal order, and word recognition. *Cognition and Brain Theory,* 7, 263-283.

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.

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

Kutas, M., & Federmeier, K.D. (In Press). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). Annual Review of Psychology.

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

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

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

McClelland, J.L., Rumelhart, D.E., & Hinton, G.E. (1986). The appeal of Parallel Distributed Processing In D.E. Rumelhart, J.L. McClelland & The PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition Volume 1: Foundations.* Cambridge: MIT Press.

O'Reilly, R.C. (1996). 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., 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.

Plaut, D.C., McClelland, J.L. (2010). Locating object knowledge in the brain: A critique of Bowers's (2009) attempt to revive the grandmotehr cell hypothesis. *Psychological Review,* 117, 284-290.

Van Berkum, J.J. A. (2008). Understanding Sentences in Context: What Brain Waves Can Tell Us. *Current Directions in Psychological Science,* 17, 376-380.

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.