Applying the dynamics of post-synaptic potentials to individual units in simulation of temporally extended ERP reading data

Sarah Laszlo (slaszlo@binghamton.edu)
Department of Psychology and Program in Linguistics, 4400 Vestal Parkway East
Binghamton, NY 13903 USA

Blair C. Armstrong (b.armstrong@bcbl.edu)
Basque Center on Cognition, Brain and Language, Paseo Mikeletegi 69, 2nd Floor
20009 DONOSTIA SPAIN

Abstract

In prior work, we have demonstrated that attention to the neural implementation of cognitive function is critical in creating models capable of simulating the physiological traces of those functions (e.g., Event-Related Potentials; ERPs). Here, we extend our Parallel Distributed Processing (PDP) model of ERP data elicited during the reading of single word forms to the simplest more temporally extended phenomenon: the ERP repetition effect. Simulations demonstrate that reproducing the dynamics of the ERP repetition effect can be accomplished by imposing the temporal envelope of post-synaptic potentials on individual units in the model.

Keywords: Parallel Distributed Processing; Event-Related Potentials; N400; Visual Word Recognition; Neural Computation

Introduction

When PDP models were first introduced in the 1980s, part of the reason for their popularity was that they allowed the simulation of cognitive function with a computational architecture that was thematically similar to that employed by real neurons. In particular, the activation of a computational unit in a PDP model is determined by weighted summation of excitatory and inhibitory input-- similar to the manner in which the potential of a neuron is determined. However, especially in the domain of word reading, the neural metaphor introduced in the 1980s has made relatively little progress since that time. Instead of focusing on improving the neural metaphor, work has largely focused on increasing the number and sophistication of cognitive tasks that can be reproduced (e.g., Harm & Seidenberg, 2004; Perry, Ziegler, & Zorzi, 2007).

This situation is unfortunate for several reasons, two of which are particularly relevant to the present research. First, the incorporation of neural constraints in PDP models, in domains besides reading, has inspired significant theoretical progress. As a representative example, consider the manner in which models implementing the details of impaired dopaminergic gating in schizophrenia have been important in outlining a unified account of the widespread cognitive impairments observed in that dysfunction (e.g., Braver, Barch, & Cohen, 1999). As we attempt to demonstrate here, similar improvements in understanding could potentially be made in the domain of visual word recognition through models implementing relevant features of neural computation.

Second, though there is substantial disagreement between modeling groups about fundamental theoretical constructs (e.g., distributed versus local representation, importance of learned behavior, importance of computational homogeneity; see Seidenberg & Plaut, 2006, for review), there is surprising agreement from many adherents of PDP models, dual-route models, and even Bayesian models, that improvement could be made to models of visual word recognition (and cognitive models more generally) by incorporating more neural constraint (Harm & Seidenberg, 2004; Perry, et al., 2007; Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010). This agreement comes at a time when there exists a similar agreement that greater computational specificity is required in theories introduced to unify a voluminous ERP reading literature (e.g., Barber & Kutas, 2007; Van Berkum, 2008; Laszlo & Federmeier, 2011).

The ERP Model

The ERP Model (Laszlo & Plaut, 2012) improves contact between computational models of visual word recognition and the neural implementation of cognitive function in two principle ways. First, the ERP model's fundamental purpose is to simulate ERP waveforms, which are direct measurements of the activity of cortical neurons. This departs from traditional reading models, which instead focus on simulation of behavioral data. In particular, the ERP model simulates key effects on the N400 ERP component. The N400 is thought to represent the obligatory access of semantics in response to the presentation of an orthographic word form (for review, see Kutas & Federmeier, 2011). This process has been explicitly couched in computational terms concordant with the PDP framework, such as

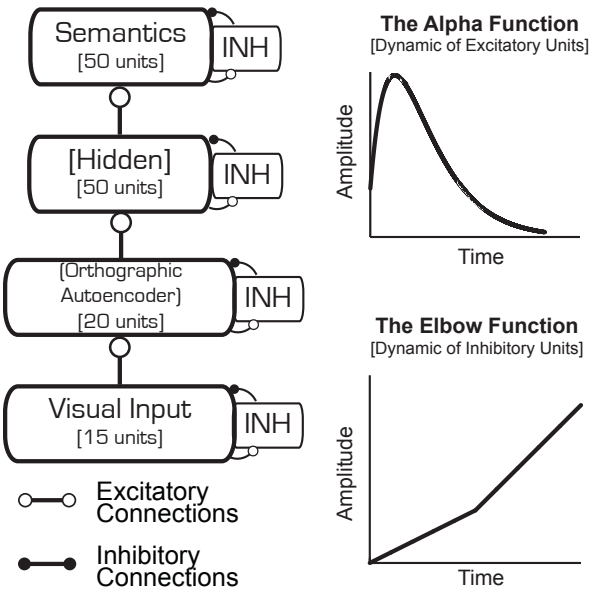

Figure 1: [Left] Architecture of the ERP model. INH stands for "inhibitory". [Right] Temporal dynamics of excitatory and inhibitory units.

parallelism and distributed representation (Laszlo & Federmeier, 2011). The ERP model has demonstrated that PDP architecture can produce the critical effects on the N400 that led to its being considered the product of PDP architecture in the first place, such as a lack of sensitivity to lexicality as compared with a much larger effect of orthographic neighhorhood size (Laszlo & Plaut, 2012).

Second, we have demonstrated that successful simulation of N400 component effects requires implementation of an important constraining characteristic of neural computation: the separation of excitation and inhibition (Laszlo & Plaut, 2012). In the ERP model, individual units have excitatory or inhibitory connections, never both. Further, inhibitory connections in the model are range-restricted, in that inhibitory connections are present only *within* a level of representation, never *between,* just as inhibitory neural projections are typically restricted to within a cortical area (this implementation is thematically similar to that in the TRACE model). Between-level connections in the ERP model are always excitatory. In addition to being range-restricted, inhibitory units in the ERP model are out-numbered by excitatory units: only one inhibitory unit is present at each level of representation. Finally, in the cortex, some populations of inhibitory units respond more quickly than others to input. In the model, this differential time course is simulated on the inhibitory units by means of the multi-linear "elbow" activation function, which produces unit activations that approximate the sum of "fast" and "slow" inhibitory sub-populations. Figure 1 displays the architecture of the ERP model and the activation dynamics for excitatory and inhibitory units. Outside of the neural constraints just described, the ERP model is a typical PDP model that follows in the tradition of PDP word recognition models that have preceded it (most recently Harm & Seidenberg, 2004). That is, its task is to associate a distributed pattern of orthographic input with a distributed pattern of semantic output, through non-linear (sigmoidal) transformation over several banks of hidden units. It accomplishes this task by acquiring connection weights over a training period of supervised learning with the back-propagation through time algorithm.

ERP Repetition Effects

The ERP model successfully simulates important component effects elicited when participants read an unconnected list of text. This type of reading material, of course, does not resemble realistic reading material in numerous respects. Most importantly for the current research, realistic text comprehension pervasively relies on context for interpretation of individual word forms. Thus, to extend the ERP model's relevance to the processes involved in reading more realistic material, it is important to extend its sensitivity to context. The simplest type of context, and a type that produces robust modulations of the N400, is the immediate repetition of a word form (e.g., DOG DOG). This simple form of context requires that the processing of word, in a minimal fashion, be dependent on what has come before it, and is thus a reasonable first step in making the bridge between simulating the response to isolated items and simulating the response to items embedded in context.

Figure 2 displays canonical ERPs elicited when words (DOG), acronyms (DVD), pseudowords (GORK), and illegal strings of letters (XFQ) are repeated. Repetition effects on the N400 are characterized by a positivity in response to a 2nd presentation, regardless of item type. The classic explanation of N400 repetition effects is that when an item is repeated in a short period of time (~10 seconds), its semantic features are still somewhat active from the prior presentation. Consequently, fewer-- unspecified-- resources need be devoted to activating the same features a second time, resulting in a reduced N400. This interpretation has been essentially unchallenged since its formation (Rugg, 1985), but, as we will see, the model will suggest a subtly different account.

ERP repetition effects are prevalent enough in not only the reading literature, but also the memory and perception literatures, that their mechanics have been considered in computational models before (Huber,

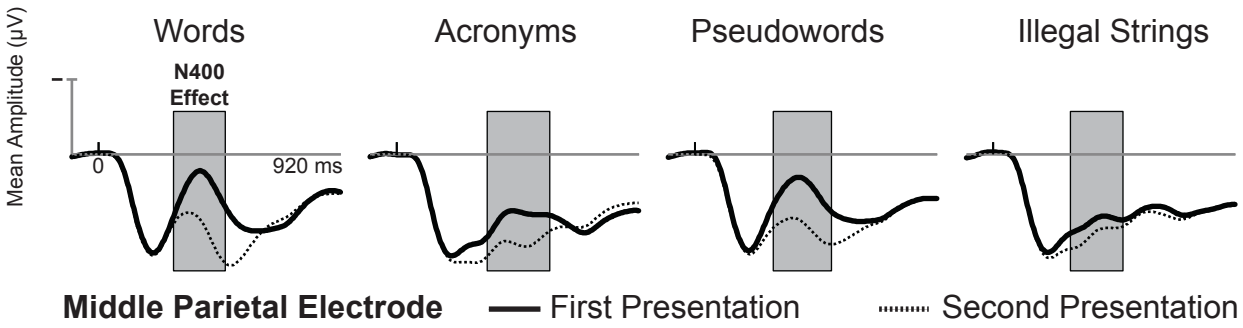

Figure 2: Grand averaged ERPs elicited in response to first and second presentations of words, acronyms, pseudowords, and illegal strings, over the middle parietal electrode. The classic N400 repetition effect—reduced N400s for repeated items—is boxed. Note: negative is plotted upwards by convention.

Tian, Curran, O'Reilly, & Woroch, 2008). This work, however, focused on early (i.e., pre-N400) repetition effects. An implemented computational account of N400 repetition effects, in contrast, is to our knowledge not present in the literature, and is a goal of the present simulations.

Unit Fatigue, Post-Synaptic Potentials, and the Alpha Function

In the model, N400 activity is linked to mean activation in the semantic level of representation. Thus, in order to effect a simulated reduced N400 in response to a repeated item, less activity must occur in semantics in the model when an item is repeated than when it is presented for the first time. In particular, specific units must become less active in response to an input when they have recently been active than when they have not; in other words, individual units must have the capacity to become selectively *fatigued*. Importantly, this fatigue must occur at the level of individual units-- not across the entire semantic level of representation-- because units that have NOT recently been active must be free to activate to their maximum level (e.g., when a new item is presented instead of a repetition).

Thus, the desired dynamic for individual units in the model in the context of item repetition is one where an initial activation peak (in response to the first item in a pair) is followed by a subsequent decline in activation. Interestingly, this dynamic profile is similar to that of post-synaptic potentials (PSPs), as simulated in neural computation with the *alpha function:*

$$V = \alpha t e^{\frac{-t}{T}} \qquad (1)$$

Where V is a measure of membrane potential, α is a scaling parameter that determines the maximum value of V, t is the number of time steps since the unit became active, and T is a free parameter that determines the time step at which V peaks (see David, Kiebel, Harrison, Mattout, Kilner, & Friston, 2006). Figure 1 displays the shape of the alpha function.

Thus, in neural computation, PSPs are simulated with a function that resembles that desired for simulation of repetition effects. This is especially interesting in light of the fact that the source of the ERP signal is cortical post-synaptic potentials. Independent observations about 1) the dynamics of the function needed to implement repetition effects and 2) the source of ERPs thus converge to suggest a method for simulating ERP repetition effects: constraint of unit activation in the model with the alpha function.

As inhibitory units in the model are already constrained with the elbow function, to allow them to simulate the response of fast and slow inhibitory populations, we confine application of the alpha function to excitatory units. We aimed to determine whether imposing this profile would enable the model to simulate ERP repetition effects.

Simulations

The architecture of the model is displayed in Figure 1, and is *identical* to that used in Laszlo & Plaut (2012), with the exception that, now, excitatory unit activation is constrained by the alpha function. To understand how this is accomplished, think of the value of the alpha function at a particular time step as a scaling parameter. In simulations, the parameter α (see Equation 1) was set such that the permitted values of V fell in [0,1]. Thus, when a unit activation is multiplied by V, that multiplication results in that unit's activation being *scaled* by V. When the alpha function is in its peak state, at $t = T$, V is 1, so multiplying unit activation by V does not change the original unit activation. However, when the alpha function is in its fatigued state, when $t > T$, $V < 1$, such that multiplying unit activations by V reduces those activations, effecting unit fatigue.

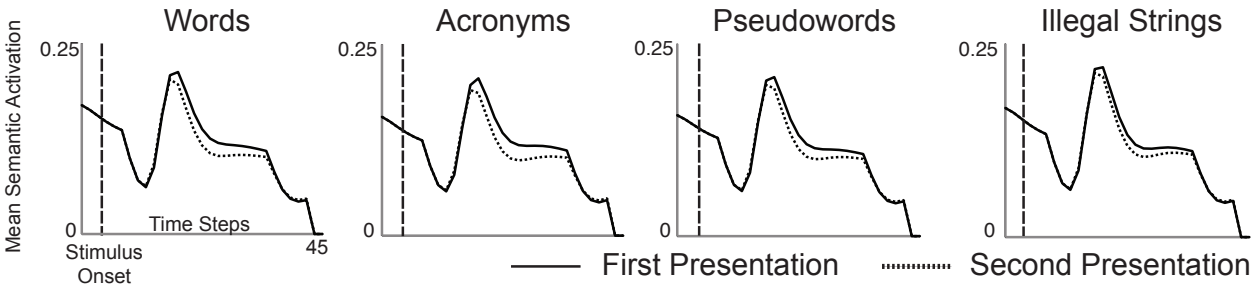

Figure 3: Simulated ERPs elicited in response to repeated and non-repeated presentations of words, acronyms, pseudowords, and illegal strings. The dashed y-axis indicates stimulus onset. All units in the model data are arbitrary. In the simulated ERPs, as in the real ERPs, all item types produce reduced semantic activation when an item is repeated as compared to when it is not.

In the cortex, of course, not all neurons generate PSPs in response to all inputs. Thus, some neurons become fatigued in response to particular inputs, and some do not. In order to implement fatigue that mirrors the cortical situation, units in the model progress along the alpha function at different rates. Specifically, *t* for purposes of calculating *V* is *not* simply the total number of time steps that have elapsed since the presentation of the input. Instead, *V* is calculated separately for each unit. In these by-unit calculations, *t* is incremented *not* with every time step in the model, *but only* when a unit's activation on the prior time step exceeded a threshold. This threshold is a fixed parameter in the model. The result of this method for determining *t* is that only units that respond to a particular input become fatigued. Units that do not respond to a particular input do not become activated above threshold, and therefore do not become fatigued.

Training

Weights in the model were initialized to small, random values. The orthographic autoencoder was then trained via back-propagation through time for 20000 epochs to reproduce orthographic inputs on an identical output layer. Then, with the weights in the autoencoder and all inhibitory weights fixed, the remainder of the network was trained for 15000 epochs to associate input orthographies with output semantics. Each training pattern was presented for 16 time steps. Training items consisted of 62 words and 15 acronyms. Importantly, the entire network's activation was reset to its initial level after each item during training, meaning that each input during training was isolated from others. Thus, the model received *no* training on repeated items. The model's output dynamics in response to repeated items must therefore be an emergent characteristic of its architecture-- newly implemented to simulate PSPs-- when extended to these novel input scenarios, *not* simply the result of training it on the desired response to repetitions.

Testing

The trained network was presented with input pairs either of the form AA (repetitions) or AB (non-repetitions). Each item of the pair was presented for 16 time steps, with a single time step of blank input between each item of the pair. In testing, the network was *not* re-initialized between items in a pair (but was re-initialized between pairs). In non-repetitions, the B item was always of the same lexical type as the A item (i.e., words were followed by words, etc.).

In addition to trained items, the network was tested on repetitions and non-repetitions of pseudowords (85) and illegal strings (279)-- these comprised all *possible* nonwords in the model's orthography. The nonwords provide a particularly hard test for the model, since they were not presented to the model during training. When presented with nonword pairs, in order to, correctly, produce reduced activation on repetition but not non-repetition trials, the model must produce dynamics it has never been trained on in response to items it has never been exposed to.

ERPs

Target ERPs for simulation were drawn from the single-item ERP corpus (for details, see Laszlo & Federmeier, 2011). Briefly, it includes responses from participants who passively read an unconnected list including 75 each of words, pseudowords, acronyms, and illegal strings-- all of which repeated once-- while EEG was recorded. Figure 2 displays the target phenomenon for simulation: N400 amplitude is reduced on second presentation for all item types.

Results

ERPs

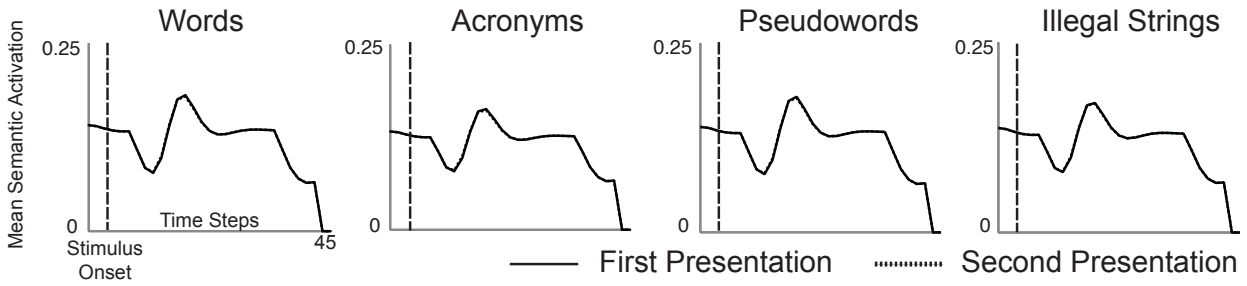

Figure 4: Simulated ERPs elicited in response to repeated and non-repeated items in a model in which the alpha function is not applied. Simulated waveforms are essentially identical across presentations in these simulations, which is why only a single wave trace is visible in the figure: the second trace is directly beneath the first. Unlike the ERPs of the alpha function model, ERPs from this simulation do not display repetition effects for any item type.

Grand-averaged ERPs were computed over the middle parietal electrode site for each item type (words, pseudowords, acronyms, and illegal strings) on each presentation (first and second). N400 peak latency was measured from 250-450 ms; N400 mean amplitude was then measured according to the full width at half max (FWHM) of that peak. This resulted in quantification of N400 mean amplitude over the 350-450 ms window. Using FWHM to determine the window of measurement allows for better consistency in measurements taken from real and simulated ERPs, as temporal units in the simulated ERPs are arbitrary (i.e., have no meaningful counterpart in milliseconds), but nevertheless have a peak and a FWHM of that peak.

The impact of repetition was assessed by analyzing the mean amplitude data for each item type using linear mixed effect regression, with item as a random factor and item type as a fixed factor. Markov Chain Monte Carlo sampling was used to generate *p*-values. These analyses replicated the standard finding: N400 mean amplitudes were reduced for all item types (all $ps < 0.0003$).

Simulations

Simulated ERPs were generated by averaging semantic activation for each time step in the model for the second item in each item pair; the time series of those averages across time steps is the simulated ERP. Figure 3 presents simulated ERPs for first and second presentations of each item type. As is evident from the Figure, simulated ERP amplitudes were reduced for each item type. Simulated N400 (sN400) peak latency was measured as simply the latency of the most positive peak in the simulated ERPs; since N400 activity is linked to mean semantic activation in the model, the peak of mean semantic activation in the model is transparently the peak of the sN400. Mean amplitude of the sN400 was then measured according to the FWHM of that peak, in analogy with measurement of the N400. Analysis identical to that described for the human ERPs revealed a substantial sN400 amplitude reduction for all item types (all $p < 0.005$).

To assess the degree to which the alpha function was responsible for the simulated repetition effects, we conducted a second simulation in which the only modification was the removal of the alpha function (essentially, this model was a replication of Laszlo & Plaut, 2012). In what follows, we will refer to this simulation as the No-Alpha simulation, and the original simulation as the Alpha simulation. Figure 4 displays results of the No-Alpha simulation. As is evident in the Figure, the No-Alpha model did not exhibit a sN400 repetition effect, in contrast with both the empirical data and the Alpha simulation. Numerically, the difference between first and second presentation sN400 mean amplitude was not different than 0 to 5 degrees of decimal precision for any item type.

Discussion

Our goal was to extend the original ERP model from being insensitive to context to being sensitive to the minimal context of whether an item has been repeated. We aimed to achieve this by improving the neural realism of the model. This improvement took the form of imposing the fatigue dynamic of PSPs on individual units in the model. The choice of this particular dynamic was motivated both by the empirical need to identify a fatiguing dynamic as well as the observation that the source of the ERP signal is cortical PSPs. Results indicated that, even when presented with a situation never encountered in training (item pairs) and items never encountered in training (pseudowords, illegal strings), a variant of the ERP model implementing unit fatigue reproduced the standard pattern observed in ERP studies: namely, that repeated orthographic items elicit reduced N400s. Importantly, reduced sN400s in response to repetition were not obtained in a version of the model without unit fatigue.

These results support the general conclusion that improving the neural realism of PDP models is a

strategy that can greatly extend the type of phenomena such models are able to explain. More importantly, however, this data provides a potential explicit mechanistic explanation of ERP repetition effects that subtly differs from that typically offered in the literature. As already discussed, the classic explanation of N400 repetition effects is that, when an item is first encountered, it invokes access of its associated semantics (or, in the case of nonwords, the semantics of visually similar items). Then, when the same item is repeated, there is less lexical-semantic processing required to re-activate the pre-activated semantics, resulting in a reduced N400 (see Rugg, 1985).

The source of N400 repetition effects in the model, in contrast, is not pre-activation of semantic features-- as is visible in Figure 3, network activity drops back almost to zero between items in a pair, before the onset of the simulated N400. Instead, semantic activity is reduced due to the fatigue of individual semantic units. While the traditional view of N400 repetition effects is based on unspecified principles of cognitive resource, the simulations suggest a view based on explicit mechanistic principles of the underlying neural system.

More exploration-- both empirical and computational-- of fatigue as an explanation of repetition effects is clearly needed: for example, it has been demonstrated in the ERP literature that additional repetitions of word forms (i.e., third, fourth, or more presentations) do not further diminish the N400 response (Young & Rugg, 2007), and it is not clear that the ERP model would exhibit this pattern. Similarly, in the present simulations words were considered a monolithic group, but it is well known that N400 repetition effects are strongly influenced by lexical factors such as word frequency (e.g., Young & Rugg, 2007), and it is again not clear that the ERP model would respond similarly. Thus, although the current work suggests an interesting alternative explanation of N400 repetition effects, based on realistic neural mechanisms and processing dynamics, clearly there is significant additional work to be done to explore this explanation further. The explicit simulation implemented here is hoped to provide a foundation for this future work.

Acknowledgments

The authors acknowledge M. Monk and the members of Binghamton University Modeling Meeting—especially K. Kurtz—for insightful discussion. This research was supported by the Research Foundation of the State University of New York and the Basque Center for Brain, Language, and Cognition.

References

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

Braver. T.S., Barch, D.M., & Cohen, J.D. (1999). Cognition and Control in Schizophrenia: A Computational Model of Dopamine and Prefrontal Function. *Biological Psychiatry*, 46, 312-328.

David, O., Kiebel, S.J., Harrison, L.M., Mattout, J., Kilner, J.M., & Friston, K.J. (2006). Dynamic Causal Modeling of Evoked Responses in EEG and MEG. *Neuroimage*, 30, 1255-72.

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.

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.

Huber, D.E., Tian, X., Curran, T., O'Reilly, R.C., & Woroch, B. (2008). The Dynamics of Integration and Separation: ERP, MEG, and Neural Network Studies of Immediate Repetition Effects. *Journal of Experimental Psychology: Human Perception and Performance*, 34, 1389-1416.

Kutas, M. and 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.

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

Laszlo, S., & Plaut, D.C. (2012). A neurally plausible parallel distributed processing model of event-related potential word reading data. *Brain and Language*, 120, 271-281.

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.

Rugg, M.D. (1985). The effects of semantic priming and word repetition on event-related potentials. *Psychophysiology*, 22, 642-647.

Seidenberg, M.S., & Plaut, D.C. (2006). Progress in understanding word reading: Data fitting versus theory building. In S. Andrews (Ed.), *From Inkmarks to Ideas: Current Issues in Lexical Processing* (pp. 25-49). Psychology Press: Hove, UK.

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

Young, M. P., & Rugg, M.D. (2007). Word frequency and multiple repetition as determinants of the modulation of event-related potentials in a semantic classification task. *Psychophysiology*, 29, 664-676.