A PDP Model for Capturing N400 Effects in Early L2 Learners during Bilingual Word Reading Tasks

Sepideh Sadeghi (sepideh.sadeghi @tufts.edu), Matthias Scheutz (matthias.scheutz @tufts.edu)
Department of Computer Science, Tufts University Medford, MA 02155 USA

He Pu (he.pu@tufts.edu), Phillip J. Holcomb (pholcomb@tufts.edu), Katherine J. Midgley (kj.midgley@tufts.edu)
Department of Psychology, Tufts University
Medford, MA 02155 USA

Abstract

Parallel Distributed Processing (PDP) models have been widely used for modeling cognitive tasks where accuracy or reaction time were the dependent performance measures. However, only few PDP models have attempted to model more brain-related data like event related potentials (ERPs). In this paper, we take a step towards using ERP data for model fitting by proposing a PDP model, which can successfully replicate various known ERP effects. Specifically, we introduce a PDP-equivalent of the N400 ERP measure and apply it to a simple PDP model of early bilingual word acquisition as bilingual word acquisition tasks provide several well-established N400 effects that can be used for model validation. We then analyze the dynamics of the network to show why and how the network can capture each of the targeted N400 effects. Furthermore, we qualitatively compare model-generated and empirical N400 peak values for L2 words.

Keywords: PDP, bilingualism, L2 word acquisition, event related potential (ERP), N400

Introduction

In a recent paper, Laszlo & Plaut (2012) proposed a way to capture N400 ERP word reading data in a parallel distributed processing (PDP) connectionist network whose architecture was based on two neurally plausible characteristics: neurons can either be excitatory or inhibitory, but not both, and inhibitory connections can only occur within layers, but not between (as the range of inhibitory connections in the brain is shorter than that of excitatory connections). Given these two constraints, the model generated cycle-based time-course data that reflected the temporal evolution of the N400 response, replicating the "orthographic neighborhood size" effect that words with larger orthographic neighborhood size elicit larger N400s compared to words with smaller neighborhood size. However, it is currently unclear whether this model would also capture various other known N400 word effects such as those obtained in the context of bilingual word processing.

In this paper, we propose a PDP architecture for a PDP model of bilingual word processing, which can successfully capture several known N400 effects in early bilingual word processing, including the "orthographic neighborhood size" effect in addition to other known effects such as the "pseudoword effect".

Background

Two important aspects of any bilingual processing model are the representations of lexical items in the first (L1) and second (L2) language and their requisite connections to concepts. Research on word processing during the early stages of L2 acquisition has revealed important constraints about storage and processing of conceptual and lexical information in the bilingual brain. Studies using speeded translation tasks, for example, show L2 learners are faster to translate from L2 to L1 (e.g., translating *tenedor* to *fork* in native English learners of Spanish) than from L1 to L2 (translating *fork* to *tenedor*) (e.g., Kroll &Stewart, 1994). These behavioral results indicate that adult bilinguals appear to associate new L2 words with their L1 translation equivalents in order to facilitate semantic access to these new words.

This bootstrapping of L2 into the already established L1 language system involves an asymmetrical representation of the two languages, accounted for in Kroll & Stewart's Revised Hierarchical Model (RHM) (depicted in Figure 1).

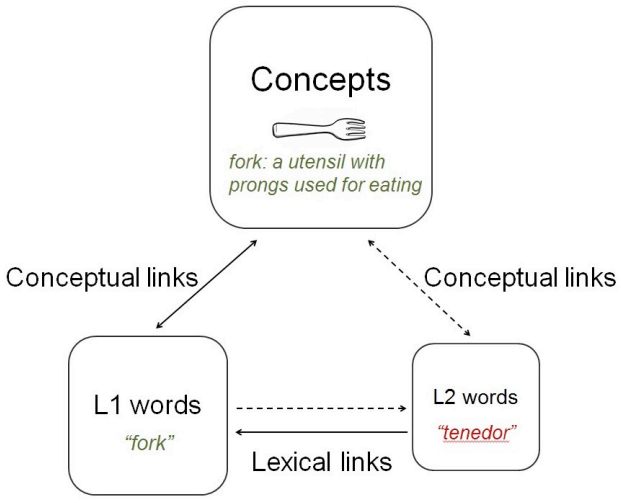

Figure 1: The RHM (Kroll & Stewart, 1994). Solid lines indicate strong connections and dashed lines indicate weak connections.

The RHM assumes a separate lexicon for L1 and L2 with orthographic and phonological representations, each of which is connected to a single amodal conceptual store. In early second language learners, the L1 lexicon is assumed to be much larger than the L2 lexicon and evidence from picture naming tasks in bilinguals suggests that the strength of the links between the two lexicons and the conceptual store are also asymmetrical, with L1 having stronger connections to semantics than does L2 (e.g., Kroll & Peck, 1998). Both the lexical level asymmetry and the concept-to-lexicon asymmetry between L1 and L2 are modeled in the RHM by disproportionately weighted links (see Figure 1). Adhering to the behavioral data, the link from the L2 lexicon to the L1 lexicon is much stronger than the link from L1 to L2, just as the link between the L1 lexicon and conceptual store is much stronger than the link between the L2 lexicon and the conceptual store.

However, behavioral data is often insufficient for distinguishing between different processing mechanisms. Hence, electrophysiological measures such as event-related potentials (ERPs) with their fine-grained temporal resolution can uncover particular neural activity elicited during language tasks that might only be associated with a particular class of model architectures. In particular, the N400, which is a negative-going centroparietally distributed ERP component peaking around 400ms after stimulus onset, has been shown to index lexico-semantic integration during word processing. Hence, it provides a robust measure of changes in processing activity in the brain as language learning takes place and can thus be used to flesh out conceptual proposals like the RHM in computational architectures such as the PDP connectionist models. We will, in particular, focus on four aspects of monolingual and bilingual word processing for which N400 effects have been reported in the literature: **(A1)** L1/L2 words versus L1/L2 pseudowords (i.e., pronounceable L1/L2 non-words that adhere to the orthographic rules of L1/L2); **(A2)** L1/L2 word repetition effects; **(A3)** variations in L1/L2 word neighborhood size; and **(A4)** L1 vs. early L2 word processing differences.

Regarding **(A1)**, it is well-known that L1 pseudowords elicit larger N400s than L1 words (e.g., Holcomb & Neville, 1990). Moreover, L2 learners showed larger N400s to L2 pseudowords than to L2 words after only 14 hours of classroom learning, mimicking L1 pseudoword effects (e.g., McLaughlin et al., 2004; however, note that McLaughlin and colleagues did not find any behavioral evidence of L2 words and pseudoword discrimination, thus supporting the use of ERPs over behavioral measures for adjudicating model architectures).

Regarding **(A2)**, repeated words reliably elicit smaller N400 amplitudes than their first presentation (e.g., Rugg, 1985). This attenuation of the N400 reflects the increased ease of lexico-semantic integration upon the second and subsequent presentations of a word (possibly due to residual activation of the lexical item and/or facilitatory feedback from the activated concept).

Regarding **(A3)**, words with large numbers of orthographic neighbors (e.g., words that differ from the target by only one letter) elicit larger N400s than words with smaller neighborhood size (e.g., Holcomb et al., 2002). Notably, the effect occurs within as well as across languages, i.e., L1 influencing L2 and vice versa (Midgley et al., 2008).

And finally **(A4)**, N400s can be used as a measure of how closely L2 processing matches that of L1 processing. For example, Midgley and colleagues found that both English-French and French-English bilinguals who had intermediate L2 experience displayed smaller N400s to L2 words than to L1 words (2009). Balanced bilinguals did not show any N400 differences between L1 and L2 word processing. This result might be in part explained by **(A3)**. Given that the L1 lexicon contains more word forms than the L2 lexicon, L1 words generally have larger neighborhood sizes than L2 words. The larger neighborhood sizes of L1 items in comparison to L2 items may contribute to larger N400 amplitudes for L1 words over L2 words.

Model Description

We start with four hypotheses, **(H1)** through **(H4)**, about the possible principles responsible for each corresponding N400 effect (i.e., **(A1)**, **(A2)**, **(A3)**, and **(A4)**) in the context of a RHM-like PDP architecture and then add connections within and between layers of the network based on the hypothesized mechanisms.

Hypotheses:
(H1) Pseudowords have no word-level representations and thus no connections to concept nodes or nodes within the lexical layer.

(H2) Concept nodes keep a residual activation between repeated word presentations and can thus be activated faster in subsequent presentations of the same word compared to the first presentation.

(H3) Lexical inputs with more orthographic neighbors should activate more concepts early on. This should lead to increased competition among concepts and thus to reduced overall activations later on, which can be facilitated via inhibitory connections in the concept layer.

(H4) After some training (when fairly strong, direct L2 lexical-to-concept connections are in place), L2 words should elicit a larger initial target concept activation than L1 words. This can be accomplished via L2-to-L1 word connections that are stronger than those from L1-to-L2 words.

Based on the RHM framework, we developed a PDP model with bidirectional excitatory lexical-to-concept connections, top-down inhibitory concept-to-lexical connections and inhibitory concept-to-concept connections (see Figure 2). As in the (Laszlo & Plaut, 2012) model, we

use IAC units (with standard parameter values for *min*=-.2, *max*=1, and *rest*=-.1 activation levels as well as *decay rate*=.1). For simplicity, we limit input words to 5 letters, thus requiring 5 clusters of 26 input letters per word (for the English alphabet). All letters in each cluster *i* have excitatory connections to words that contain them in the *i*-th slot and inhibitory connections to all words with a different letter in the *i*-th position.

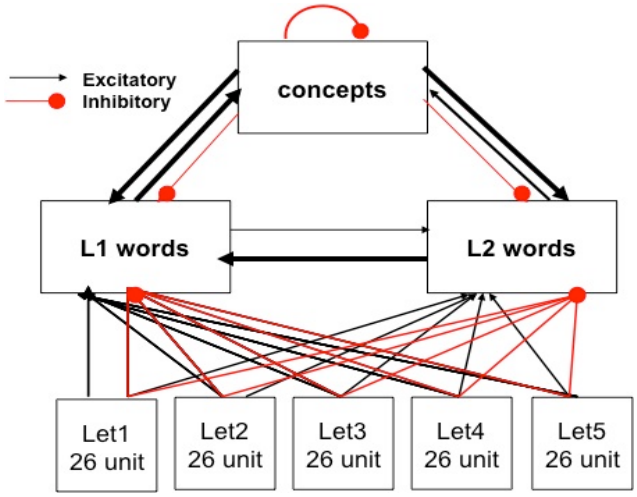

Figure 2: Model architecture. The thickness of links indicates the strength of connections.

L1 versus L2. To account for larger L1 vs. L2 word neighborhoods, we include more L1 words with a larger neighborhood size in the model compared to L2 words.

Pseudowords versus words. Pseudowords have no representation at the lexical or semantic layer.

Repetition. We model repetition effects by performing the following sequence *r* times: input word *i* is presented for *n* cycles (where *n* should be large enough to allow the N400 signal, to be defined below, to reach its peak). Then the input is removed and the network is updated for *d* cycles to let all node activations decay, after which point the whole process is repeated, but without resetting any activation values. We thus have three critical modeling parameters that need to be set appropriately: *r*, *n*, and *d*.

Filtering word length artifacts. Assuming that each constituent letter contributes equally to a word's activation level, all connection weights from each letter in a word have the same strength. However, because words have different lengths, the overall incoming activation would be different if we were to use the same connection weights for all letter-to-word connections as longer words would get a higher activation than shorter words, everything else being equal. To avoid this effect, we scale the letter-to- word connection weight *c* by the length $|W|$ of the word W: $w_{L,W} = c/|W|$. We also needed to make sure that the input letters corresponding to a given target lexical item will only activate the orthographic neighbors and not the other words that differ from the target word in more than one letter. In order to do so, we made the strength of inhibitory and excitatory letter-to-word connections the same, so that if a word is different from the target word in more than one letter (for four-letter words), it receives zero or less than zero netinput from the letter nodes. In addition, none of the five-letter/three-letter words were similar to a four-letter word in 3 or more slots.

Definition of PDP N400 Measure

Based on the semantic interpretation of the N400 signal (Laszlo & Fedemeier, 2011), we define the network-equivalent of the N400 as the magnitude of overall activation change (differential) in positively activated (potential) concept nodes (potential). Specifically, we calculate the sum of all positive concept activations at each cycle and compute the change between two consecutive cycles as the N400 (the discrete equivalent to the derivative of the potential given by the summed concept node activations).

Experimental Bilingual ERP Data

We collected ERP measures from 14 native English speakers who were enrolled in a first semester "Introductory Spanish" class at Tufts University (9 females, mean age 18.4). Participants viewed Spanish words (e.g., HOLA, GATO) and Spanish pseudowords (e.g., SERO, AGOL) one at a time as part of a lexical decision task. The Spanish words were a set of non-cognates taken from the textbook used in class. Factors of length, English bi-gram frequency, and English neighborhood size were balanced between the words and pseudowords used in the study. Averaged ERPs were computed for all word and pseudoword stimuli for each participant at 29 scalp sites. Single item ERPs were formed by averaging to time-locked stimuli across participants. The mean amplitude averaged across a subset of centroparietal electrodes (including: Cz, Pz, C3, CP5, CP1, P3, C4, CP6, CP2, P4) between 300-500 ms was used to quantify the N400 effect. The mean amplitude between 300-500 ms was used to quantify the N400 effect. Additionally, N400 measures for single items were calculated using the mean amplitude between item-specific temporal windows, ranging from 250ms to 500ms.

Modeling Results

We selected a subset of 14 four-letter L2 words from all L2 words used in the ERP experiment and included all their L1 translations as well as their L1 neighbors to be able to account for the cross-language orthographic neighborhood size effects. Since some of the L1 words were 5 letters in length, we included 5 clusters of letters in the model.

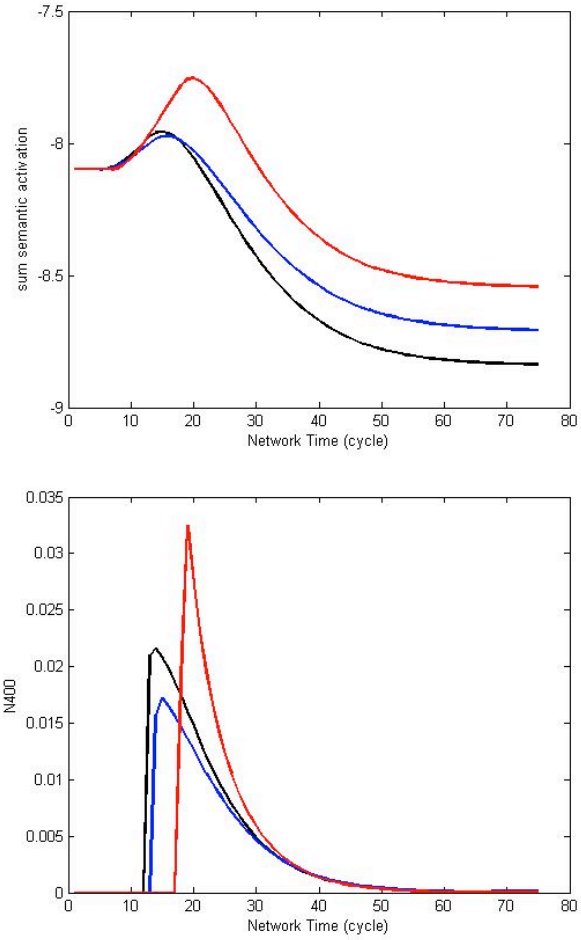

Figure 3: Sum of semantic activation (top row) and the N400 amplitude (bottom row), over 75 update cycles in response to three words: "son" (L1 word in black), "azul" (L2 word in blue), and "sero" (L2 pseudoword in red).

Figure 3 shows the shape of the N400 signal generated by the model along with the time-course of the summed concept nodes' activations during the whole word exposure. Note that the change in total concept activation is proportional to the maximum value of the N400 generated.

The right column in Figure 3 reveals three distinct phases in the dynamic of the overall semantic activation in our network: (a) charge (positive overall change), (b) discharge (negative overall change), and (c) stabilization. Furthermore, since inhibitory connections only originate from concept nodes, any significant flow of inhibition can only come after an initial flow of activation, i.e., until concept nodes have reached sufficiently strong activations.

Charge. The activation of the target concept and concepts associated with orthographic neighbors or its associated word-level node initially start to increase, followed by the feedback from excitatory and inhibitory connections to word-level nodes causing the activation of the target word to gradually increase and the activations of its orthographic neighbors to decrease.

Discharge. The overall semantic activation decreases as a result of inhibition exerted by significantly activated concept nodes.

Stabilization. Eventually, the overall activation levels of the network stabilize.

We searched for values for the various connections that would allow the model to capture the N400 effects: concept-to-L1=(.6,-2), concept-to-L2=(.8,-.2), concept-to-concept=(0,-.6), L1-to-concept=1, L2-to-concept=.8, L1-to-L2=.1, L2-to-L1=1, letter-to-(3letterWord)=(.8,-.8), letter-to-(4letterWord)=(.6,-.6), letter-to-(5letterWord)=(.48,-.48) (the first element of each tuple is the excitatory weight value between related items, and the second element is the inhibitory weight value between the unrelated items).

For all simulations, we took the maximum peak value as the measure for comparing N400 signals to the empirical data. Furthermore, since several factors can influence the N400 value, we investigated only one factor at a time while keeping the others fixed.

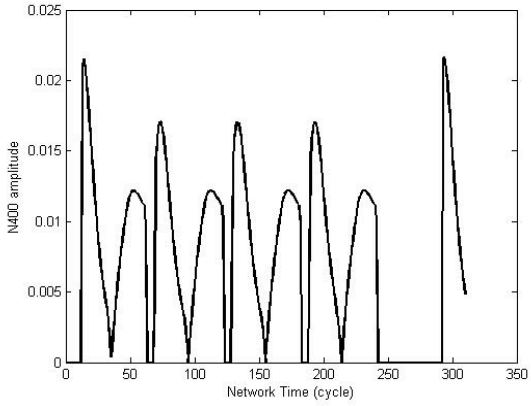

Figure 4: N400 data for repetitions of "son" using first: *r*=3, *n*=30, *d*=30, second: *r*=1, *n*=30, *d*=70, and then *n*=30 (see text for details).

Figure 4 shows that the model replicates the repetition effect **(A2)**, i.e., maximum N400 values (peaks) after the first exposure are all smaller than the first peak.

Figure 5 shows that the model is able to replicate the neighborhood size effect regardless of lexical type: L1, L2, and pseudowords.

Figure 6 shows the replication of **(A4)** – in all cases – and the replication of **(A1)** – in all cases except for (a) and (b). Furthermore, Figure 4 suggests that the replication of **(A1)** and **(A4)** is dependent on neighborhood size: as the neighborhood size increases, the network replicates **(A1)** more strongly, while showing weaker replication of **(A4)**. The network best replicates **(A4)** for L2 words of *nSize*=0.

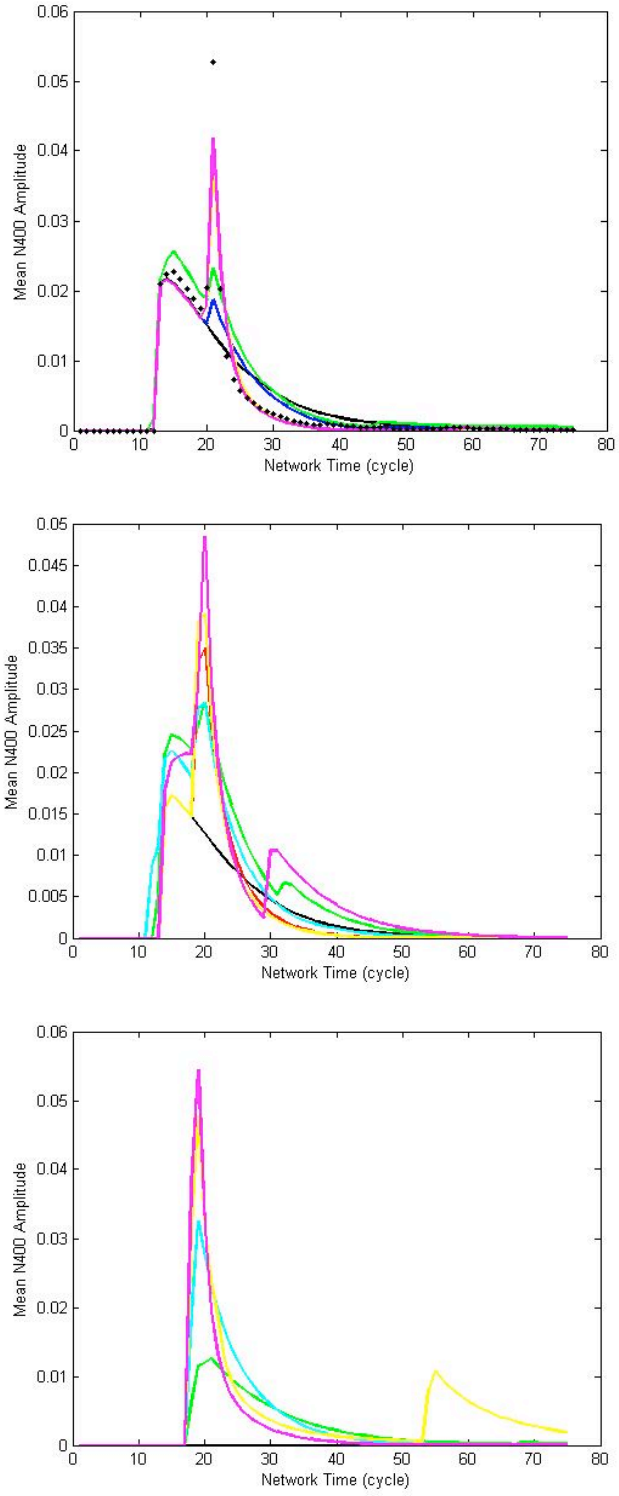

Figure 5: Neighborhood size effects within three categories: in order L1, L2, Pseudowords, shown by mean N400s of words with *n* orthographic neighbors: 0=black, 1=blue, 2=green, 3=cyan, 4=red, 5=yellow, 6=magenta, & 10=black stars.

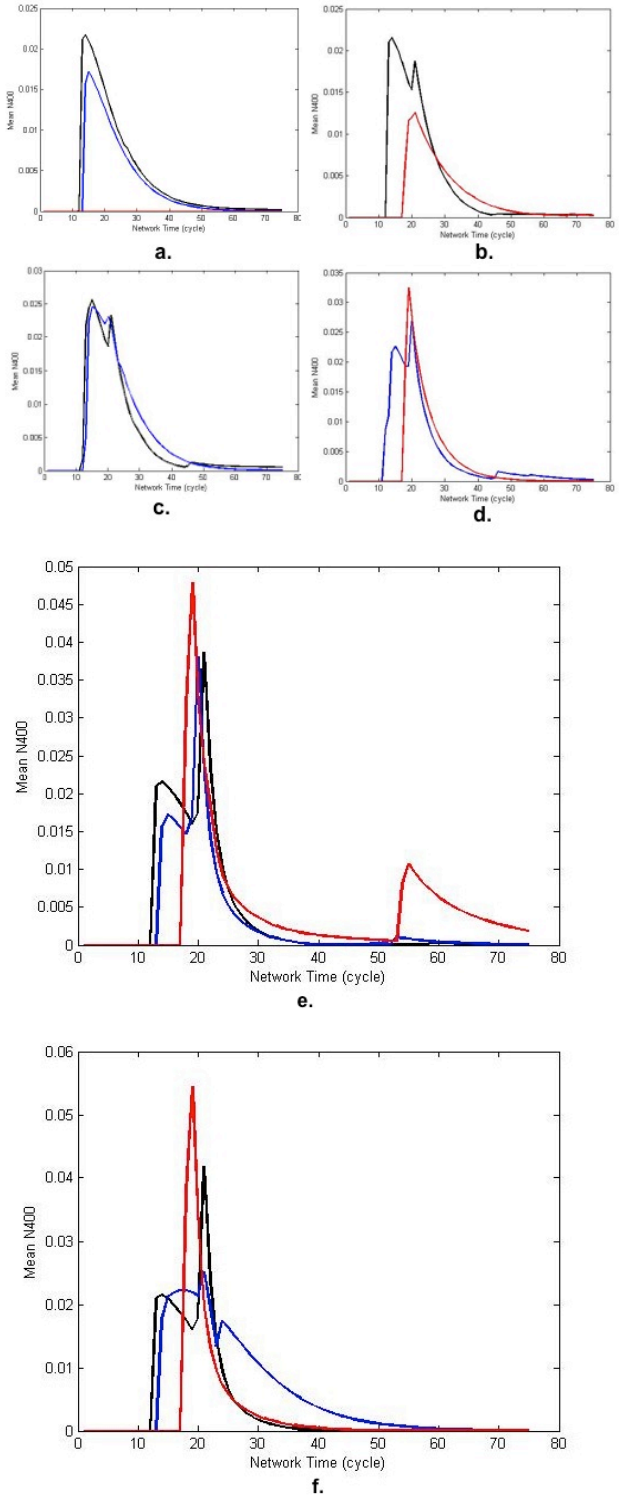

Figure 6: Mean N400 signals for words sharing the same neighborhood size (*nSize*): a) *nSize*=0, b) *nSize*=1, c) *nSize*=2, d) *nSize*=3, e) *nSize*=5, f) *nSize*=6, in three categories: L1 words in black, L2 words in blue, and L2 pseudowords in red. Note that there was no L2 word of *nSize*=1 , no pseudoword of *nSize*=2, and no L1 word of *nSize*=3.

Note that the correlation value (*corr*=.2135) between the maximum N400 values (for L2 words) generated by the model and those collected in the experiments shows that the model does not yet quantitatively fit the empirically obtained ERP values, despite qualitatively replicating ERP effects.

Discussion

The model succeeded in capturing qualitatively all four ERP effects. Furthermore, the results confirm that the **(A1)** and **(A4)** effects are dependent on neighborhood size as suggested in (Midgley et al, 2008 and Holcomb & Neville, 1990). However, the model allows for a different explanation from that of Midgley et al. who hypothesized that the overall lower N400 for L2 words compared to L1 words might be caused by the smaller neighborhood size of L2 words compared to L1 words, everything else being equal. Specifically, the model shows that this difference can also be obtained with identical neighborhood sizes based on the generally higher initial activation induced at the target concept in response to an L2 input word (compared to that of an L1 input word). This higher initial activation tends to suppress the other concept nodes, thus leading to an overall lower ERP and thus lower N400. Hence, it is likely that both neighborhood size and difference of initial target concept activation via L1 or L2 words contribute to the smaller N400 for L2 words (compared to L1 words).

Note that all simulation results where obtained by considering N400 peak values only, but other measures are certainly possible (e.g., the integral of the N400 signal over the 300-500msec time frame or the average value over the same period). This is left for future work.

Conclusion

We have developed a PDP model based on Kroll & Stewart's Revised Hierarchical Model (RHM) of bilingual word processing and tested it against well-established N400 effects. The model succeeded in qualitatively replicating language, neighborhood size, pseudoword, and repetition effects. However, the model did not quite replicate the N400 results from our empirical experiments, as shown by a fairly low correlation between the ERPs of the model and empirical data. Future work will focus on exploring the model's parameter space to determine if better model fits are possible with the given model architecture. In addition, we will investigate simpler model architectures and the extent to which they may succeed in replicating some of the N400 effects. We will also investigate alternative definitions of N400 (e.g., including the lexical level activations) as well as exploring the use of average N400 amplitudes rather than peak values.

References

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.

Holcomb, P. J & Neville, H. J. (1990). Auditory and visual semantic priming in lexical decision: A comparison using event-related brain potentials. *Language and Cognitive Processes*, *5*, 281-312.

Kroll, J. F., & Stewart, E. (1994). Category interferences in translation and picture naming: Evidence for asymmetric connection between bilingual memory representation. *Journal of Memory and Language*, *33*, 149-174.

Kroll, J. F., & Peck, A. (1998). Competing activation across a bilingual's two languages: Evidence from picture naming. *Proceedings of the 43rd Annual Meeting of the International Linguistic Association*. New York University, NY.

Laszlo, S., & Plaut, D.C. (2011). Simulating event-related potential reading data in a neurally plausible parallel distributed processing model. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

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.

McLaughlin, J., Osterhout, L., & Kim, A. (2004). Neural correlates of second-language word learning: Minimal instruction produces rapid change. *Nature Neuroscience*, *7*, 703–704.

Midgley, K.J., Holcomb P.J., van Heuven, W. J. B., & Grainger, J. (2008). An Electrophysiological investigation of cross-language effects of orthographic neighborhood. *Brain Research*, *1246*, 123-135.

Midgley, K. J., Holcomb, P. J., & Grainger, J. (2009). Language effects in second language learners and proficient bilinguals investigated with event-related potentials. *Journal of Neurolinguistics*, *22*, 281-300.

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

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