A Neural Field Model of Word Repetition Effects in Early Time-Course ERPs in Spoken Word Perception

Andrew P. Valenti (andrew.valenti@tufts.edu)
Human-Robot Interaction Laboratory, Tufts University, 200 Boston Ave.
Medford, MA 02155 USA

Michael Brady (michael.brady@tufts.edu)
Human-Robot Interaction Laboratory, Tufts University, 200 Boston Ave.
Medford, MA 02155 USA

Matthias J. Scheutz (matthias.scheutz@tufts.edu)
Human-Robot Interaction Laboratory, Tufts University, 200 Boston Ave.
Medford, MA 02155 USA

Phillip J. Holcomb (pholcomb@tufts.edu)
Department of Psychology, Tufts University, 490 Boston Ave.
Medford, MA 02155 USA

He Pu (he.pu@tufts.edu)
Department of Psychology, Tufts University, 490 Boston Ave.
Medford, MA 02155 USA

Abstract

Previous attempts at modeling the neuro-cognitive mechanisms underlying word processing have used connectionist approaches, but none has modeled spoken word architectures as the input is presented in real-time. Hence, such models rely on the ingenuity of the modeler to establish a mapping of real-time stimulus to the model's input which may not preserve processing that happens during each time step. We present a neural field model which successfully replicates the effect of immediate auditory repetition of monosyllabic words and fits it to a component of a well-studied mechanism for analyzing language processing, the event-related potential (ERP). This represents a new modeling approach to studying the neuro-cognitive processes, one that is based on the bottom-up interaction of real-time sensory information with higher-level categories of cognitive processing.

Keywords: dynamic neural fields; event-related potential (ERP); spoken word perception; mental workload; computational modeling; word repetition

Introduction

By "spoken word perception", we mean the cognitive processes that entail the sensory intake of an acoustic waveform until the words contained in it are identified. Some early connectionist models of speech perception processes were driven by research in generalized automatic speech recognition and have shown, for example, that a good deal of phonemic information is present in the auditory signal and can be extracted from the statistical generalization of the model. Among the best-known models of speech perception is TRACE (McClelland & Elman, 1986) which has modeled several lexical effects (e.g., phonemic restoration in a noisy environment) and the time-course of word recognition. TRACE has been criticized for its biologically unrealistic handling of time and the lack of a learning mechanism (Protopapas, 1999). As a result, models were developed (Elman, 1990; Norris, 1995) which represent time through cyclical, recurring connections from one state to an earlier state in the network. One popular method by which learning is incorporated in these networks is through a gradient decent regression using backpropagation.

While these models can account for many aspects of how humans comprehend spoken and written words, none of these architectures model speech perception using real-time, human input. We present a neural field model with an efficient learning mechanism which dynamically responds to the spoken word process as it unfolds over time. A neural field sits in an equilibrium state waiting for a pattern it has tuned itself to detect, and this detection takes the form of a perturbation. Learning associates the equilibrium state of a field with its environment. Primary fields tune themselves to fall into systematic equilibrium states in response to combinations of sensory input. Deeper-processing, secondary neural fields are then enabled to tune themselves in response to their environments once primary fields have settled into predictable behaviors. With experience, the network forms representations as each neural field systematically responds to its environment through time.

Word Repetition Effects and ERPs

An event-related potential (ERP) is an electrical voltage associated with an event such as a stimulus or response. ERPs are believed to reflect the summation of post-synaptic potentials occurring in many thousands of neurons. The time course of ERPs in auditory processing can be traced starting from stimulus onset and continuing for approximately 800 ms. Our study focused on a particular ERP known as the P200 (P2) which occurs in the interval from 145 ms to 225 ms after

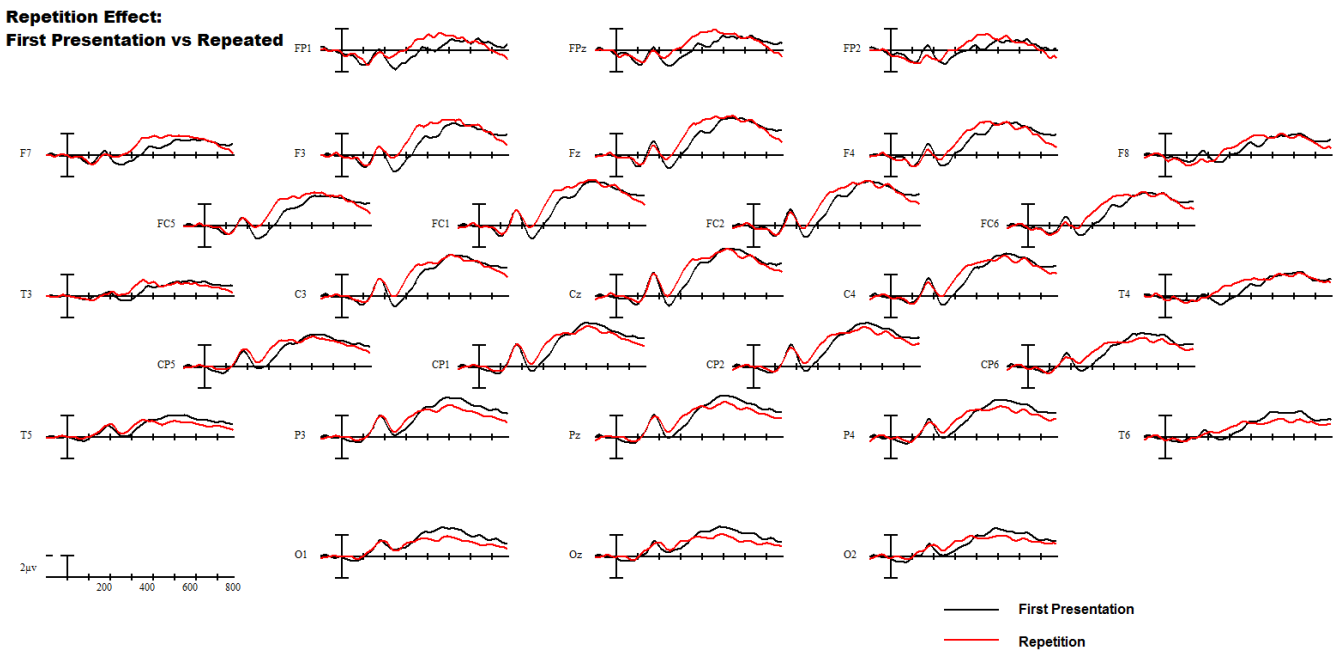

Figure 1: ERP repetition effects, seen in the difference between the first presentation (black line) or a word and the immediate repetition (red line) of that word

stimulus onset and is classically associated with top-down attention processes on early sensory processing (Hillyard & Anllo-Vento, 1998). Of particular interest, the P2 has also been associated with a word repetition effect (Luck, 2014; Molfese, Key, Maguire, Dove, & Molfese, 2005) where the P2 showed a reduced positivity (i.e., a larger negativity) to primed versus unprimed targets. Word repetition is frequently used as an investigative tool in psycholinguistic and memory research. It is a simple empirical procedure which demonstrates that subjects are usually faster in their response to the second presentation of words than the first; such responses may be captured via reaction time (RT) measures across a variety of experimental paradigms such as lexical decision or semantic categorization.

Prior research in which participants read short texts containing repeated words has found three distinct ERP components to be sensitive to repetition: a positive component peaking around 200 ms post-stimulus, a negative component at 400 ms (N400) and a later positivity (van Petten, Kutas, Kluender, Mitchnier, & McIsaac, 1991). However, van Petten et al. (1991) note that the early P2 repetition effect has not been consistently found in other studies, at times appearing with an opposite polarity. Due to the paucity of research using real-time speech signals and the conflicting early results cited, it appears that the processes which control this early component are not well-understood. Among the research questions that remain open are to what extent does deeper lexical processing and explicit memory influence the word repetition effect and what particular cognitive processes elicit this effect? While this paper did not set out to explore these questions in depth, we address some of them in the context of our results.

Human Experiments and ERP Data

Empirical ERP Data

We collected ERP data from 12 Native English speakers from Tufts University (mean age 19.6, 7 male), of which 2 were excluded due to excessive ocular artifacts. All participants self-reported as monolingual and right-handed (Oldfield, 1971), with normal or corrected-to-normal vision/hearing and normal neurological profile. Participants provided written informed consent and were monetarily compensated, as approved by the Tufts University Institutional Review Board.

Materials and Design

During ERP recording, participants completed a dual-task paradigm with a primary task of playing a video game (i.e., "Breakout": breaking pre-arranged blocks by bouncing a ball from a controllable paddle) and a secondary task of listening to words through a set of headphones. The dual-task paradigm was important for our ERP modeling task because we attempted to reduce any explicit episodic memory effect so that we could focus on more implicit repetition primary effects by introducing the primary task of playing a video game. For the primary task, we utilized a JavaScript variant of Breakout. Three game levels were chosen based on pilot results, indicating them to be similar in difficulty. For the secondary task, a female experimenter recorded 300

monosyllabic English words to be used in stimuli generation. These 300 words were split into two lists (of 150 words each) matched for psycholinguistic properties (e.g., bigram frequency, length, phonological and orthographic frequency, familiarity, and concreteness). An additional list was created from the two split lists (half from each) so that a total of three lists of 150 words were created. From each of the 3 lists, 50 of the 150 words were randomly selected to be repeated so that each list contained a total of 200 words. None of the repeated words were redundant across lists.

EEG Recording

Participants engaged in the dual-task paradigm in a dark, sound-attenuated room while their EEG was recorded using a 29-channel electrode cap. Loose electrodes recorded from 1) below the left eye (LE) to monitor for blinks and vertical eye movements, 2) at the right temple (HE) to monitor for horizontal eye movements, and 3) behind each mastoid (left: A1, right: A2) for referencing (A1) and monitoring differential mastoid activity (A2). Electrode impedances were kept under 5 k for all scalp electrodes, 10 k for both eye electrodes, and 2 k for both mastoid electrodes. We sampled the EEG at 200Hz while an SA Bioamplifier (SA Instruments, San Diego, CA) amplified the signal with bandpass of 0.01 and 40 Hz.

Experimental Results

Averaged ERPs were formed for each spoken word (using -100 and 0 ms baseline) after artifact rejection (15.67% of the trials were rejected due to ocular artifacts) and collapsed into conditions (first presentation or repeated) for comparison. The ERPs were then low-pass filtered at 15 Hz. Individual participant ERPs were then averaged into a grandmean of 10 participants, allowing for the analysis of overall auditory language processing effects. Of particular interest is the repetition effect on particular ERP components such as the P2 (van Petten & Kutas, 1991; Rugg, 1987) with an anterior scalp distribution, sensitive to lexical processing and implicated in word recognition processes (Dambacher, Kliegl, Hofmann, & Jacobs, 2006). Such repetition effects manifest in the form of attenuated amplitudes to repeated items compared to their first presentation, reflecting the ease of processing for the former relative to the latter. Results indicate the presence of a P2 repetition effect, seen clearly in anterior electrodes between 200 and 400 ms (Figure 1).

Model Description

We modeled a single layer of the hierarchical process generally regarded to represent the architecture of speech perception (Grossberg, 2005; McClelland & Elman, 1986; Norris, 1995). In Figure 2, the model architecture consists of (1) a vector of *auditory input nodes*, (2) a vector of *category nodes*, (3) a grid of processing units called a *neural field*, and (4) three fully connected sets of weights to be trained called *adaptive filters*. The field processing units are reciprocally connected to each other through non-adjustable

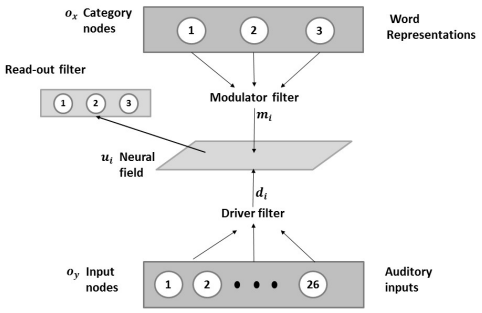

Figure 2: Neural field training. The training vector at the word representation layer develops an input signal $s = m_i$ through the modulator filter to each processing unit u_i in the neural field as a random sound exemplar of the same training vector category is played to the input nodes.

weighted connections using an on-center, off-surround "Mexican hat" distance function (Brady, 2014). The input nodes carry sensory information which is refreshed with new data at each time step. This input is passed through a "driver" filter to develop a bottom-up input signal to the field. The category nodes carry persistent labeling information which is passed through a "modulator" filter to provide a top-down input signal to the field. The labeling information is also used as the training target for a "read-out" filter.

A neural field in our model is a "sheet" of processing units. If given no input and random initial conditions, all units of the field are guaranteed to quickly fall into a stable equilibrium state with respect to each other such that the entire field may be considered to fall into an equilibrium. Different equilibrium states of the field are associated with different input patterns. The field is updated once every 10 ms (i.e., a time step) using Equation 1 which computes the change in its activation. This general equation and its variations are widely used in dynamical systems models, (e.g., Amari, 1977; Beer, 2000; Brady, 2014; Grossberg, 2005; Hopfield, 1982; Schöner & Spencer, 2015).

$$\dot{u_i} = -u_i + s_i + h + n + \sum_j \lambda(i,j) \cdot \sigma(u_j) \qquad (1)$$

The change in activation of a unit, u_i at a given time step is computed as the sum of influence to the unit at that time step minus the activation of the unit from the previous time step. Influence to a unit at a time step comes from an input signal, s_i, the field's slightly negative bias, h, a noise term, n, and from other units within the field. Influence from other units within the field is computed to be the sum of the squashed activations of neighboring units multiplied through corresponding within-field connection weights w. A stepwise squashing function, σ, is used such that only units with non-negative activations can influence their neighborhoods. Within-field con-

nection weights are specified as on-center off-surround by a Mexican hat weighting function, $\lambda(D)$. Input to the function D is the Euclidean distance between two units, u_i and u_j; the output of the function specifies their connection strength.

Neural Field Learning

We implemented a learning mechanism in which the driver and modulator filters are trained together that works as follows. The filter weights are initialized with random values which are then updated across training cycles. A training cycle consists of iterations in which the neural field is initialized with random unit activations simulating the passage of time between learning patterns. Then, a training vector is used to generate an input signal s_i through the filters to each unit of the neural field using Equation 2, and a random sound exemplar of the same category as the training vector is played to the input nodes as time unfolds. In our experiment, the training vector represents a monosyllabic word. Here, o_y is the activation of a category node, o_x is the activation of an input node, and g_1, g_2, g_3 are gain terms; $\dot{d}_i$ is the change in activation of the driver signal, $\overline{u}_i$ is the running average of the unit being updated, and $\overline{m}_i$ is the running average of the modulator signal to a unit.

$$s_i = g_1|\dot{d}_i| \cdot (g_2\overline{m}_i - g_3\overline{u}_i) \tag{2}$$

$$\overline{m}_i = \sum_y w_{iy} \cdot o_y$$

$$\dot{d}_i = \sum_x w_{ix} \cdot o_x$$

The weights of the modulator and driver filters are adjusted following Equation 3, a variant of the delta training rule.

$$\Delta w_{ix} = \eta \cdot \bar{o}_x \cdot (\overline{u}_i - \dot{d}_i) \cdot |\dot{u}_i| \tag{3}$$

$$\Delta w_{iy} = \eta \cdot \bar{o}_y \cdot (\overline{u}_i - \overline{m}_i) \cdot |\dot{u}_i|$$

Learning proceeds as the training vector persists for the duration of the input sound as the neural field adjusts itself in response to its input, updating the modulator and driver filters at each time step. Subsequently, a new iteration begins by initializing the field to a new random state and associating the transformation of that state through time with the next input training vector (i.e., new word), and so on. A cycle is completed when all training vectors have been exposed to the model in random order, at which point a new training cycle begins.

In Equation 3, η is the learning rate, $(\overline{u}_i - \dot{d}_i)$ and $(\overline{u}_i - \overline{m}_i)$ are the errors to be minimized; cyclic training continues until the learning error is reduced to asymptote. The last term of the equation, $|\dot{u}_i|$, is an innovation which allows learning to occur only if there is a change in the target neural field and therefore important associations are maintained even as learning proceeds over time.

Neural Field Model Simulations and Results

The model's read-out filter is trained in order to evaluate how well the neural field categorizes its input. The weights of this read-out filter are updated using the "delta rule" as in Equation 4. Training vectors o_y are converted to target vectors T_y by setting the negative values of the training vectors all to zero. The generated output is notated as $\hat{o}$.

$$\Delta w_{yi} = \eta \cdot u_i \cdot (T_y - \sigma(\hat{o}_y)) \tag{4}$$

Where:

$$\hat{o}_y = \sum_i w_{yi} \cdot m_i$$

We selected a subset of five monosyllabic words from the stimuli used in the empirical experiment: "beach", "dog", "soup", "bog", and "tend". Four exemplars of each word were recorded separately by a male speaker as male voices span a lower frequency range making for easier speech processing by the model. The recordings were transformed into the 26 coefficients shown in Figure 2 and were provided as input to the model in 10 ms time steps. The model was trained on three target words from this set, "beach", "dog", and "soup". How well the model learned was measured by computing the error as the sum of the differences between "readout" vector generated as output by the model and the corresponding target word.

Modeling the ERP Measure

We chose to model the ERP as the difference between the modulator signal and the field activation. This can be thought of as analogous to error values or implicit prediction error. Implicit prediction error at multiple levels of language processing is thought to play a critical role in language comprehension (Kuperberg & Jaeger, 2015). Within probabilistic frameworks, implicit prediction error has been linked to other language-related components such as the N400 ERP (Kuperberg, 2013; Xiang & Kuperberg, 2014; Kuperberg, 2016), as well as non-linguistic ERP components (e.g., Friston, 2005; Wagongne, Changeux, & Dehane, 2005). Moreover, the N400 ERP component has recently been simulated as cross-entropy error at a semantic level within a connectionist model (Rabovsky & McRae, 2014).

The ERP at time t is computed as shown in Equation 5; m_i and u_i are each unit's modulator and field activation respectively:

$$ERP_t = \sum_i |m_i - u_i| \tag{5}$$

Modeling Results

The words from the test input were presented in the following order: "soup", "dog", "dog", "dog", "beach", "dog", "bog", "tend". The neural field was trained on "soup", "dog", and "beach"; "bog" and "tend" were novel stimuli the field was not trained on. Figure 3 shows that the model replicates the repetition effects, i.e., the maximum ERP values at a t after the first exposure of the word "dog" are all smaller than

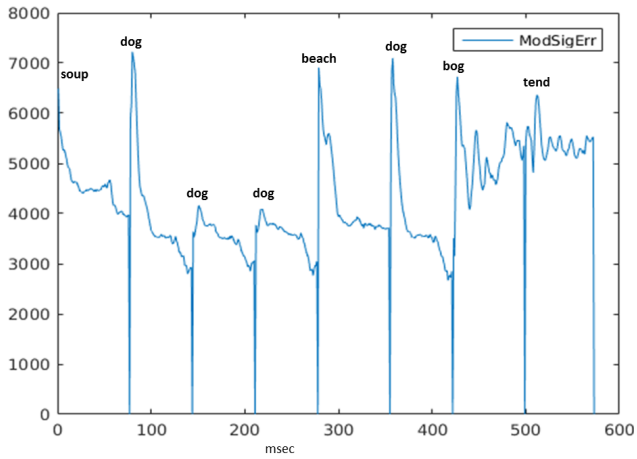

Figure 3: Modulator-Field difference for repetitions of the word "dog"

the first peak, until a different word is presented. At this time, the neural field is perturbed into a different state, releasing it from the effect. A subsequent presentation of "dog" no longer elicits a repetition effect, producing a larger peak as the field resettles into the equilibrium state for "dog". In Equation 5, the modulator signal, m_i, can be thought to "predict" the next equilibrium state the neural field u_i is likely to settle to. This suggests that a smaller amount of perturbation is required to "nudge" the settled field into a new equilibrium state upon presentation of a repeated word. The presentation of untrained, novel stimulus, i.e., "bog" and "tend", does not show the repetition effect as these words are not predicted by the modulator signal.

Table 1: Model Fitting

Interval Width	Model Proportion	ERP Data Proportion
100 ms	1.41	1.60
112 ms	1.53	1.53
120 ms	1.66	1.58
Best fit (112 ms) 144 ms - 256 ms		

We note that model does not aim to fit the polarity of the P2 ERP as what gives rise to the polarity is not well-understood and as there have been inconsistent reports on the word repetition effect as mentioned earlier (van Petten et al., 1991). Furthermore, it is the nature of ERP measurement that the interval within which a given effect is manifested varies somewhat between experimental paradigms. However, the model should fit the magnitude and the duration of the human ERP data. Thus, to compute the model fit, we looked at the ERP data intervals centered around 200 ms as this interval contains the P2 effect and computed the proportion as follows. We took the area under the ERP curve within an interval for the first presentation of the word "dog" and divided it by the identical interval contained under the repeated presentation to calculate its proportion. Referring to Figure 3, we also took the area under the ERP curve generated by the model and performed the same calculation. As shown in Table 1 we found that the 112 ms interval around 200 ms (i.e., from 145 ms to 255 ms) showed both proportions to be identical i.e., 1.53, thus demonstrating it is possible to find a good model fit to the experimental data.

Discussion

We designed our model to be a single neural field reflecting processing in the auditory cortex and hypothesized that this forms a "layer" of phonological processing. In order to provide a modulator signal, we simulated the existence of a deeper word-form layer by "clamping" the modulator signal to the three words the model was trained on (i.e., "beach", "dog", "soup") and this was fed "down" to the neural field as its modulator signal. We did not presuppose which ERP correlates would occur using only one neural field layer and did not set as a goal to identify all possible auditory effects; we were not concerned with capturing non-speech auditory processing at all.

The model succeeded in capturing the repetition effect noted in the experimental results as can be seen in Figure 1, most notably in the central scalp ERPs e.g., Cz. Figure 3 shows a diminished response to the initial presentation of the word "dog" at 75 ms with the repetition effect occurring at 150 ms and 225 ms. Note that the typical convention is to plot the ERP, with the area above the x-axis as negative and the area below as positive. Thus the model and ERP waveforms covary in amplitude and polarity with the repetition (i.e., in the model the repetition effect is "more negative" than the initial presentation).

The model demonstrated the immediate word repetition effect using a single neural field sheet, without modulator input from deeper lexical and semantic processing layers. This suggests that the ability of a single neural field layer to learn sound patterns (i.e., phonemes, monosyllabic words) alone appears to be sufficient to account for the immediate word repetition effect and the release from repetition. We believe this to be among the first computational models to match the time course of ERP events on real-world, real-time data, and the first model to do so using spoken word perception i.e., we used the same data that was presented to the experiment's participants and validated the model fit. These results suggest that our neural field approach can now be used to build additional layers and thus model later ERPs.

Conclusion

We have developed a dynamic neural field model of phonological processing of monosyllabic spoken words and compared it with a separately designed experiment which measured ERP responses of participants to spoken words. We

found a good fit between the model and the human ERP data. The model succeeded at replicating the word repetition effect showing a positive correlation with the experiment's P2 measurements. This suggests that a minimal neural field model can perform some components of auditory processing (e.g., detect immediate word repetition) and generate a correlated ERP effect. Future work will explore modeling deeper lexical and semantic processing and related mid-to-late ERP effects by connecting additional neural field layers in a hierarchy which will allow feedback from the deeper processes to affect computations at earlier layers.

Acknowledgments

Research was sponsored by the U.S. Army Natick Soldier Research, Development, and Engineering Center and was accomplished under Cooperative Agreement Number W911QY-15-2-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Natick Soldier Research, Development, and Engineering Center, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

References

Amari, S. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological cybernetics*, *27*(2), 77–87.

Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Science*, *4*(3), 91–99.

Brady, M. (2014). A bi-directional graphical model for babble-feedback learning in speech. *Procedia Computer Science*, *41*, 220–225.

Dambacher, M., Kliegl, R., Hofmann, M., & Jacobs, A. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain Research*, *1084*(1), 89–103.

Elman, J. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society London, Series B, Biological Sciences*, *360*(1456), 815–836.

Grossberg, S. (2005). Adaptive resonance theory. In L. Nadel (Ed.), *The encyclopedia of cognitive science* (1st ed.). Wiley.

Hillyard, S., & Anllo-Vento, L. (1998). Event-related brain potentials in the study of visual selective attention. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(3), 781–787.

Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities. In *Proceedings of the national academy of sciences* (Vol. 79, pp. 2554–2558).

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

Kuperberg, G. R. (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience*. doi: 10.1080/23273798.2015.1130233

Kuperberg, G. R., & Jaeger, T. F. (2015). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*. doi: 10.1080/23273798.2015.1102299

Luck, S. J. (2014). *An introduction to the event-related potential technique* (2nd ed.). Cambridge, MA: The MIT Press.

McClelland, J., & Elman, J. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86.

Molfese, D., Key, A. P. F., Maguire, M. J., Dove, G., & Molfese, V. (2005). Event-related potentials (ERPs) in speech perception. In D. Pisoni & R. Remez (Eds.), *The handbook of speech perception.* Blackwell Publishing.

Norris, D. (1995). A dynamic-net model of human speech recognition. In G. Altmann (Ed.), *Cognitive models of speech processing.* Cambridge, MA: MIT Press.

Oldfield, R. (1971). The assessment and analysis of handedness: the edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113.

Protopapas, A. (1999). Connectionist modeling of speech perception. *Psychological Bulletin*, *125*(4), 410–436.

Rabovsky, M., & McRae, K. (2014). Simulating the N400 erp component as semantic network error: insights from a feature-based connectionist attractor model of word meaning. *Cognition*, *132*, 68–89.

Rugg, M. (1987). Dissociation of semantic priming, word and non-word repetition effects by event-related potentials. *The Quarterly Journal of Experimental Psychology*, *39*(1), 123–148.

Schöner, G., & Spencer, J. (2015). *Dynamic thinking: a primer on dynamic field theory*. New York, NY: Oxford University Press.

van Petten, C., & Kutas, M. (1991). Electrophysiological evidence for the flexibility of lexical processing. In G. Simpson (Ed.), *Understanding word and sentence.* Amsterdam: North-Holland Press.

van Petten, C., Kutas, M., Kluender, R., Mitchnier, M., & McIsaac, H. (1991). Fractionating the word repetition effect with event-related potentials. *Journal of Cognitive Neuroscience*, *3*(2), 131–150.

Wagongne, C., Changeux, J.-P., & Dehane, S. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *360*(1456), 815 – 836.

Xiang, M., & Kuperberg, G. R. (2014). Reversing expectations during discourse comprehension. *Language, Cognition and Neuroscience*. doi: 10.1080/23273798.2014.995679