Does predictive processing imply predictive coding in models of spoken word recognition?

James S. Magnuson (james.magnuson@uconn.edu)

Monica Li (monica.li@uconn.edu)

Sahil Luthra (sahil.luthra@uconn.edu)

Heejo You (hee_jo.you@uconn.edu)

Rachael Steiner (rachael.steiner@uconn.edu)

Psychological Sciences & CT Institute for the Brain and Cognitive Sciences, U. Connecticut, Storrs, CT 06269-1020

Abstract

Pervasive behavioral and neural evidence for *predictive processing* has led to claims that language processing depends upon *predictive coding*. In some cases, this may reflect a conflation of terms, but predictive coding formally is a computational mechanism where only deviations from top-down expectations are passed between levels of representation. We evaluate three models' ability to simulate predictive processing and ask whether they exhibit the putative hallmark of formal predictive coding (reduced signal when input matches expectations). Of crucial interest, TRACE, an interactive activation model that does not explicitly implement prediction, exhibits both predictive processing and model-internal signal reduction. This may indicate that interactive activation is functionally equivalent or approximant to predictive coding, or that caution is warranted in interpreting neural signal reduction as diagnostic of predictive coding.

Keywords: prediction; predictive coding; language; computational modeling; neural networks

Prediction in spoken language processing

Listeners often predict upcoming information in spoken language. They anticipate upcoming phonemes based on lexical expectations (Grosjean, 1980; Allopenna, Magnuson, & Tanenhaus, 1998), and upcoming words based on lexical, syntactic, and/or discourse expectations (Altmann & Kamide, 2007; Magnuson et al., 2008; Strand et al., 2018). There is also neural evidence consistent with prediction. Indeed, many ERP studies test the magnitude and timing of responses to expectation violations, including responses that precede complete bottom-up specification. Despite difficulties replicating one classic example (Delong, Urbach, & Kutas, 2005 vs. Nieuewland et al., 2018), a large number of studies support varying degrees of prediction (for reviews, see Kuperberg & Jaeger, 2015; Hickock, 2012).

Evidence for predictive *processing* (PP) is often considered evidence for *predictive coding* (PC), and there may be instances where these terms are conflated and treated synonymously. PC, however, is a computational formalism enaling efficient coding by comparing bottom-up inputs to predictions from a top-down model and passing forward (and backward) only *deviance from prediction* (Rao & Ballard, 1999). This deviance is the novel *information*; sending bottom-up details would be redundant when predicted by higher-level expectations. Thus, formal PC predicts reduced feedforward and feedback signal when inputs conform to top-down expectations. In light of several reports of neural signal reduction when word-level expectations are met (e.g., Blank & Davis, 2016; Gagnepain, Henson, & Davis, 2012), we next consider what evidence for PP and PC implies for models of spoken word recognition (SWR).

Implications for models of spoken word recognition

First, even without considering sentence-level contexts (beyond the scope of current models), models of SWR must be able to simulate attested word level PP. Intuitively, some models might do this readily (e.g., a simple recurrent network [SRN; Elman, 1990] trained to predict the next phoneme given the current phoneme), while others may not. For example, Gagnepain et al. (2012) suggest that the interactive activation model, TRACE (McClelland & Elman, 1986), may be inconsistent with PC because they describe its primary mode as *competitive* rather than *predictive*.

Second, given growing neural evidence consistent with formal PC (reduced neural signal when expectations are confirmed vs. violated; e.g., Sohoglu & Davis, 2016) we can also ask whether a model of SWR exhibits this hallmark of PC: internal signal reduction when expectations are confirmed. This leads us to two questions for models of SWR. (1) Do models with explicit prediction (e.g., SRNs) and without explicit prediction (e.g., TRACE) simulate PP? (2) If so, do they show hallmarks of formal PC (model-internal signal reduction when expectations are confirmed)? To address these questions, we will compare three models.

Model comparisons

Our simulations are based on human experiments by Gagenpain et al. (2012). In those experiments, there were three critical stimulus types: an **Original** word (e.g., *formula*), a **Trained** nonword (e.g., *formubo*), and an **Untrained** nonword (e.g., *formuty*). In the examples, we have underlined letters corresponding to the critical phonemes. Prior to training, both Trained and Untrained nonwords differ from expectations at 1-3 phonemes from offset; this position follows the *deviation point*. The critical question is how the system responds at the phoneme(s) following the deviation point before a training phase and after. In the training,

participants get extensive exposure to the *trained* nonwords. Prior to training, Gagnepain et al. (2012) found reduced neural activity in left superior temporal gyrus following the deviation point for Original items vs. both types of nonwords. Following training (and sleep), Trained items showed the same reduction relative to Untrained items as real (Original) words. In the following sections, we examine whether each of 3 models is able to simulate PP (sensitivity to expectations at the deviation point) in simulations of this paradigm, and whether they exhibit the hallmark of PC: signal reduction when top-down lexical expectations are met.

Model 1: Predictive Cohort

Gagnepain et al. (2012) used a simple mathematical model to generate predictions for their experiment. We call their model "predictive cohort" because it simply looks up the set (cohort) of words that remain consistent with the phoneme-by-phoneme input for each item. For example, given *formula*, at position 1, all /f/-initial words are possible and the prediction for position 2 is the frequency-weighted probability distribution of each phoneme following /f/. As input progresses, probability distributions narrow. For *formula* (/formjul^/), by /u/ at position 6, very few possibilities remain (*formula, formulaic, formulation*) and all predict /l/. Gagnepain et al. (2012) derived positional prediction error for the three item types. Given a prediction of 1.0 for /l/ at position 7, *formula* would garner zero prediction error, while prediction error would be high for *formubo* and *formuty*.

The logic is that a formal PC implementation would pass back a prediction of /l/ at position 7 given the input for positions 1-6, and therefore pass forward a very weak signal given *formula,* where the prediction error is low, compared to the nonword cases. Note that prediction error is not an internal signal in this model; it is a derived term meant to stand in for computations that would occur in formal PC.

Methods

Materials We implemented predictive cohort as described by Gagnepain et al. We selected 37.6k words ≤ 12 phonemes long from the English Lexicon Project (ELP; Balota et al., 2007). Critical items were 54 Original-Trained-Untrained triples from Gagnepain et al. (mean length: 6.3 phonemes). Deviation points were 1-3 positions before offset.

Procedure We conducted two suites of simulations with all 54 x 3 = 162 items. In *pretraining*, the lexicon was restricted to 37.6k real words; thus, the Original items were words, and the Trained and Untrained items were nonwords. *Post-training*, Trained items were simply added to the lexicon, changing the positional probability distributions embedded in the lexicon (as done by Gagnepain et al.). For each simulation, we computed predicted probability distributions at each position, and calculated implied prediction error.

Results are presented in Fig. 1. Consistent with PP, the probability for Original items continues to increase beyond the deviation point at Pretraining, and probabilities also increase for Trained items Post-training. Because error is summed over all phonemes, the maximum is 2.0 (e.g., if predicted values for /l/ and /b/ were 0.8 and 0.0, but the input were 0.0 for /l/ and 1.0 for /b/, summed error would be 1.8). Error plots do *not* reflect model-internal information. Rather, error is meant to approximate what the forward signal would be *if a formal PC model were implemented.* Thus, while the predictive cohort model is able to exhibit PP, it does not inherently exhibit PC. Of course, a fully-implemented PC model would show such signal reduction.

Model 2: Simple Recurrent Network

The second model we tested was a Simple Recurrent Network (SRN; Elman, 1990). An SRN would seem likely to naturally produce PP, given that an SRN is typically trained to predict

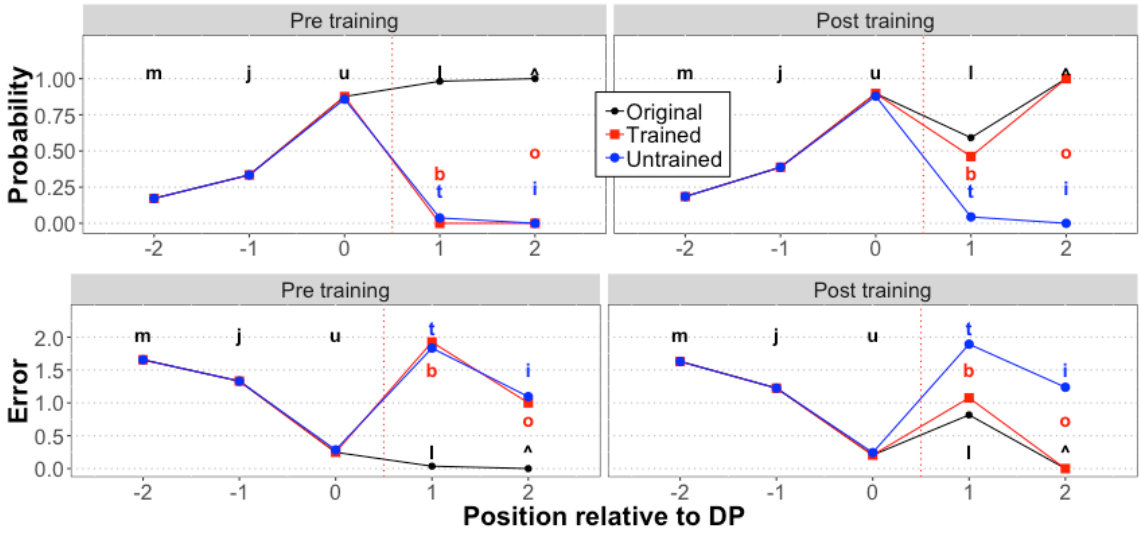

Figure 1: Predicted phoneme-by-phoneme probabilities (top) and derived errors (bottom), pre- (left) and post- (right) training for the Predictive Cohort model. The X-axis is position relative to the deviation point (allowing us to align results for all items). The dashed lines between positions 0 and 1 indicate the deviation point.

the next item in a series. We created an SRN with localist phonemic inputs (41 nodes, one for each phoneme) with forward connections to 200 hidden units with forward connections to 41 localist phonemic output nodes. The hidden nodes feed an exact copy of their states with a 1-cycle time delay to *context* nodes, which feedback to hidden nodes (providing a memory sensitive to multi-step contingencies).

Methods

Materials We used the same materials as for Model 1.

Procedure The model was presented with a continuous series of phonemes constructed by randomizing the order of the 37.6k words and presenting each phoneme-by-phoneme, without any break or indication of a word boundary. The network was trained using backpropagation of error to predict the next phoneme. At each time step, output activations were compared to the desired output pattern (1.0 for the following phoneme, 0.0 for all others). Backpropagation allows "blame" to be assigned to all connections in the network (i.e., to calculate how small changes to all weights could alter the network such that if the same input sequence were applied again, the network would come closer to the target pattern).

After approximately 2000 epochs (each epoch is 1 pass through all 37.6k words in random order), error plateaued (aggregated over small batches of words). This does not mean error rate was uniform. Rather, output activations come to resemble the probability distributions calculated by the predictive cohort model (Model 1). Thus, error is relatively high near word onset and diminishes as the input progresses.

For the *pretraining* test of the model, only the 37.6k words selected from the ELP for Model 1 were included. Because the SRN is a learning model, we were able to actually train the model on Trained items. The 54 Trained items were presented in novel random orders for 50 epochs. This number of instances was sufficient for the model to achieve Original-level accuracy with Trained items without impairing the model's ability to process items already in its lexicon.

Results are in Fig. 2, and are similar to those from Model 1, but with output activations for relevant phonemes. Error indicates the summed error over all 41 output phoneme units. Like Model 1, the SRN exhibits PP pre- and post-training in that phonemes from trained items become more probable after training. Also like Model 1, though, note that error is not a model-internal value; it is calculated externally. Model-internal signals (here, activations) *do not exhibit the reduced-signal hallmark of PC*. Instead, activations are *higher* when expectations are met (when the input sequence corresponds to a word in the lexicon). Thus, even the most intuitively predictive model of SWR one might propose (short of a formal PC model) – an SRN – does not inherently exhibit PC.

Some might disagree with this analysis, since SRNs are trained using backpropagation of error, and these error terms could be considered to be passed back through the model, even if error is typically not passed during tests and is not necessary for a trained SRN to function. We might counter that backpropogation is model-external (the procedure is not part of the network dynamics of an SRN; adjustments to weights are imposed on the network, rather than an emergent property. One might contrast this with Hebbian learning, where weight changes occur through biologically-inspired interactions among nodes. On the other hand, while backpropogation may not have a direct analog in biology, functionally-equivalent, neurally-plausible mechanisms are not far-fetched (Lillicrap & Santoro, 2019). It may be sensible, then, to consider the error signal in an SRN as a feedback signal, in which case SRNs show the PC hallmark of relative signal reduction when inputs match expectations.

Model 3: TRACE

TRACE (McClelland & Elman, 1986) is an interactive activation model: a neurally-inspired, parallel-distributed processing model with feedforward connections from inferior to superior levels (features→phonemes→words) and lateral inhibition within levels. It also has feedback from words to constituent phonemes. As mentioned earlier, TRACE may

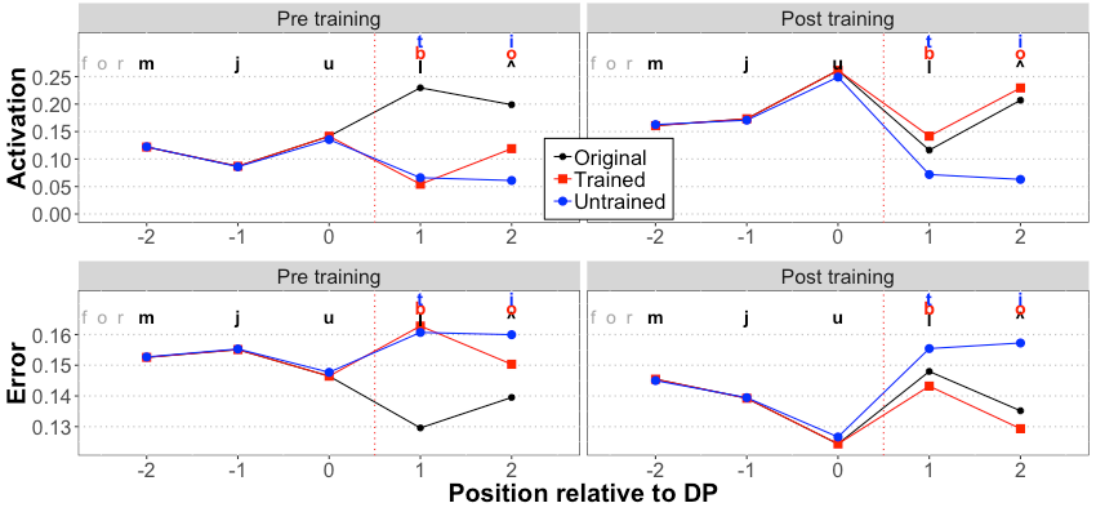

Figure 2: Phoneme-by-phoneme SRN output activations indicating how strongly the model predicted each upcoming phoneme (top) and those activations converted to error scores over time (bottom).

not seem to be a predictive model (e.g., Gagnepain et al. [2012] describe TRACE as having a primarily competitive mode of processing, dominated by lateral inhibition). However, word→phoneme feedback in TRACE provides a generative model (McClelland, 2013; Magnuson et al., 2018): as features and phonemes consistent with a specific word are presented, the lexical node for that word form sends increasingly strong feedback to *all* its constituent phonemes, including those that have not yet occurred. This allows graded pre-activation of phonemes (as a function of how strongly *expected* they are due to feedback from one or more words). But is there any possibility that TRACE exhibits the hallmark of PC – reduced signal when expectations are confirmed vs. violated?

Methods

Materials We used the original 212-word TRACE lexicon, wherein we identified 15 six-phoneme words on which to base item sets. From each set, we created 2 nonwords by changing the final two phonemes. For example, from the Original /art^st/ (*artist*) we created /art^da/ and /art^pi/.

Procedure Simulations were conducted with all 15 (set) x 3 (item type) items. We tracked activations of phonemes and words over time as well as the total amount of activation (and inhibition) flow between and within levels during each simulation. For pre-training, the lexicon consisted only of the TRACE lexicon, including the 15 Original items. For post-training, the 15 Trained items were added to the lexicon.

Results We begin by comparing lexical activations for each item type (Original, Trained, Untrained) pre- and post-training (Fig. 3). Pre-training, we see significantly weaker Original activation when input ends with final phonemes of Trained or Untrained items. Post-training, we see a decrease in Original activation given Untrained input, and a massive decrease given Trained input. This is because Trained items are now words in the lexicon; with clear input, Trained items strongly activate and inhibit their Original counterparts. The post-training panel in Fig. 3 includes a red line marked with an open red square; this indicates activation of Trained items given corresponding input. This line is directly on top of the Original line; since both items are words in the lexicon, clear corresponding input drives both similarly.

Next, consider the phoneme level (Fig. 4). Activations of phonemes one position beyond the deviation point are plotted for the Original word, as well as replaced phonemes in the case of the Trained and Untrained nonwords. In Fig. 4, we can see differences in the lines with open symbols that achieve high activation. These correspond to activations of replaced phonemes (/d/ in /art^da/ or the /p/ in /art^pi/). Pre-training, the highest activation is achieved for the phoneme in penultimate position in the Original word, thanks to support from both bottom-up input and top-down lexical feedback. There is only a slight disadvantage for the replaced phonemes; given clear bottom-up input, phonemes will be strongly activated, even in the absence of lexical support. Post-training, with Trained items added to the lexicon, the 'replaced' phoneme in a Trained item achieves nearly identical activation as a phoneme in an Original item, since both receive lexical support.

To address PP, Fig. 5 zooms in on the regions delineated with dashed squares in Fig. 4. Pre-training, the activation of the Original phoneme is higher than that for replaced phonemes beginning ~12 cycles prior to the deviation point. Phoneme activations from cycles ~18 to ~33 (just past the deviation point, indicated by the dashed vertical line) are driven nearly exclusively by top-down feedback. Bottom-up input begins to override feedback just after the deviation point. At this point, when the input has a replaced phoneme (one of the nonworeds), the activation of the Original phoneme drops, while activation of the replaced phonemes when they are actually the input (dashed lines, open symbols) jumps dramatically. Post-training, we see a lexical advantage for phonemes after the deviation point for both Original and Trained items (for Trained items, activations after the deviation point is slightly less due to a small trend for those items to have lower transitional probability in the lexicon, even when they have been added to the lexicon). In summary, training elicits clear PP: increased activation of critical phonemes prior to the deviation point.

Next, let's consider PC, which could manifest as reduced feedforward or feedback signal when expectations are confirmed; to be fully consistent with PC, both the feedforward and feedback signal would have to be reduced when expectations are met. However, the standard in many cognitive neuroscience studies is that any evidence of signal reduction is taken as evidence for PC. We therefore tracked the total amount of activation flowing between levels

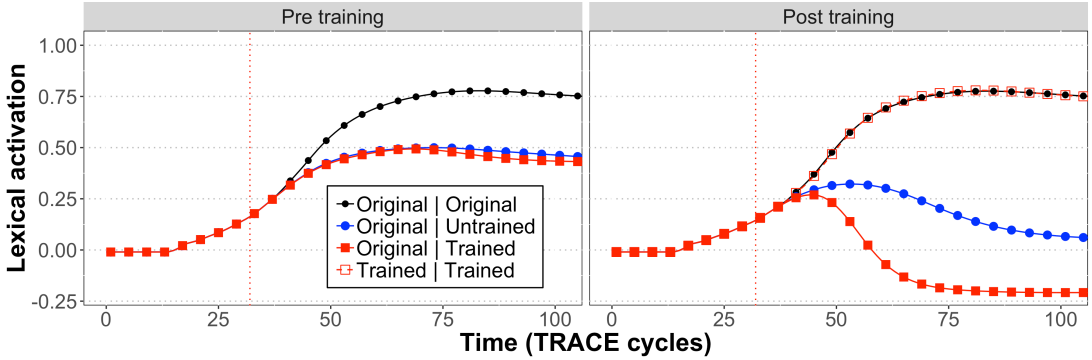

Figure 3: Lexical activations in TRACE before and after 'training'. Note that 'Trained | Trained' is only valid post-training.

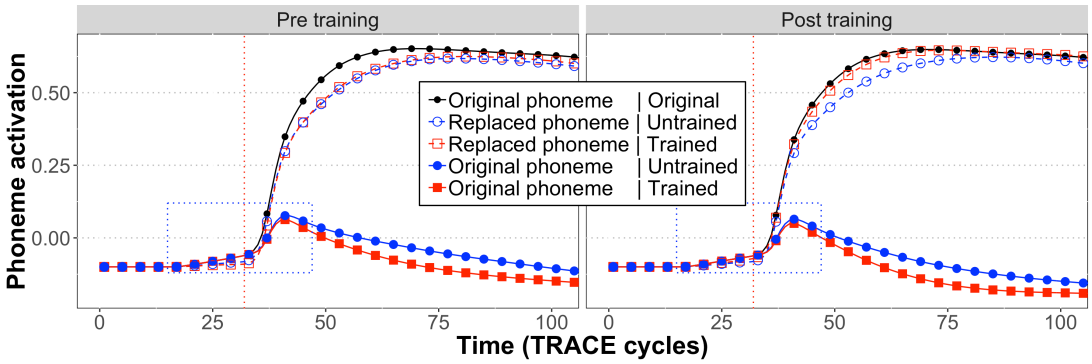

Figure 4: Activations of critical phoneme (following deviation point) in TRACE.

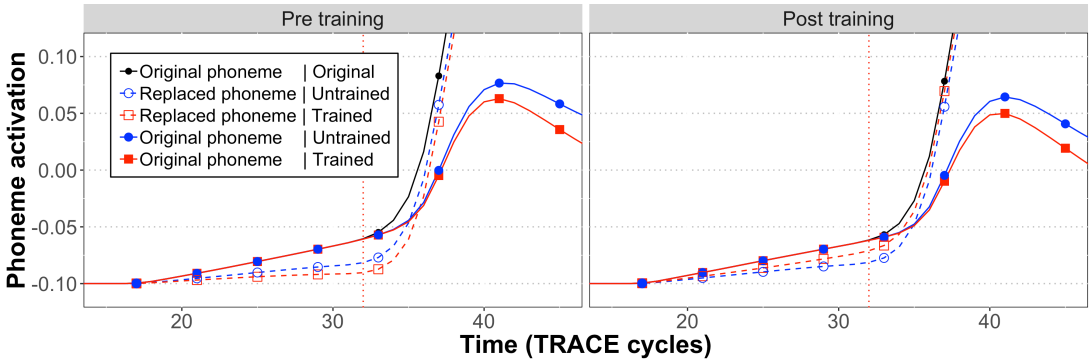

Figure 5: Zoomed view of critical time period from Figure 4.

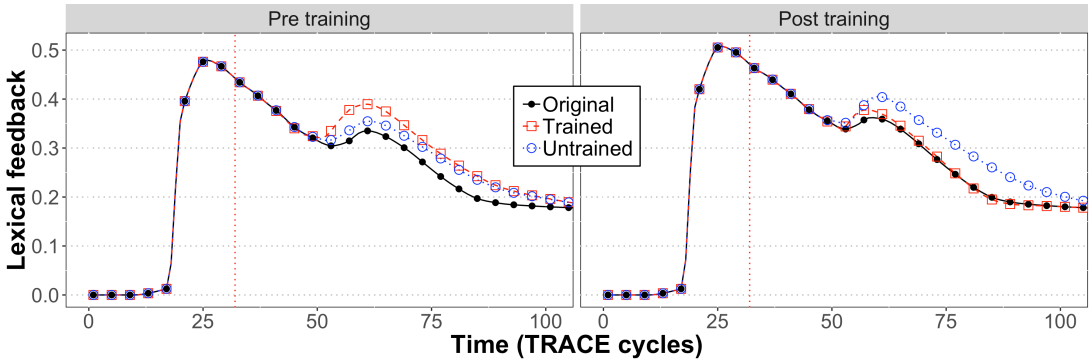

Figure 6: Total lexical feedback over time in TRACE, showing robust signal reduction when expectations are met.

(forward and backward) and within levels (lateral inhibition), looking for any signal reduction.

Two activation indices were reduced when expectations were met: word→phoneme feedback (Fig. 6) and lateral inhibition (absolute value). The latter shows virtually the same pattern as Fig. 6, but we omit it due to length constraints and challenges in interpreting a reduction in a signal with negative valence. In Fig. 6, total lexical feedback is *lowest* when expectations are met (for Original words pre- or post-training, as well as for Trained items post-training). This is because when an unexpected phoneme occurs, Original items already have strong support and continue to send substantial feedback. Additional feedback comes from words partially activated by replaced phonemes (any word unit containing the unexpected phoneme *aligned* with the unexpected phoneme[s] would get activated; e.g., a word unit for *piano* aligned at position 5 overlaps with the /pi/ of /art^pi/]). This follows from the *total* amount of feedback actually being less when one word can strongly dominate and inhibit other words; there can actually be *more* total feedback when many words are weakly activated. Thus, only TRACE, the model one might have predicted to be least likely to exhibit PC, shows a *model-internal signal reduction* often considered *diagnostic* of PC in cognitive neuroscience.

Discussion

All three models tested – predictive cohort, an SRN, and TRACE – exhibit PP. The first two showed model-internal signal *increases* when expectations were met. While these increases can be converted to predicted error, this takes place outside the current instantiation of these models (though see our earlier discussion of backpropagated error in SRNs). TRACE shows model-internal signal *reduction* when expectations are confirmed, in the form of lesser top-down

lexical→phoneme feedback.

This raises the possibility that interactive activation (as implemented in TRACE) may provide a generative model that is functionally equivalent (or functionally approximant) to a Bayesian generative model (McClelland, 2013) or even PC. Addressing this question will require the development of explicit, formal PC models of SWR based on formalisms like those introduced by Rao and Ballard (1999). This is a tall order; such a model must work on over-time inputs (if not real speech), must be validated with a moderately large lexicon (at least hundreds of words), and must be comprehensively compared to other models, such as TRACE.

There are promising starts in this direction. For example, Yildiz et al. (2013) have reported a PC model of SWR that operates on real speech. However, this model was limited to a 10-word vocabulary (names for the digits 0 to 9). Another promising example comes from Blank and Davis (2016), who implemented simple network models of SWR with lexical→phoneme feedback that was either multiplicative (as in TRACE) or subtractive (one possible interpretation of PC). Both models correctly simulated one experiment, but their subtractive feedback model correctly predicted neural signal reduction in a second experiment where the multiplicative model predicted signal increase (but with radical parameter changes required to fit the two experiments; in one, they ran models for more than 300 cycles, while for the second, they ran models for only 1 cycle). This sort of work, along with comprehensive tests of models on at least moderately large vocabularies (to verify that the models are consistent with known facts about SWR), are needed to advance understanding of the potential role for PC in SWR.

In the absence formal PC models, we must exercise caution when interpreting neural signal reduction. Though our results indicate that TRACE exhibits model-internal signal reduction, it remains an open question whether interactive activation is indeed functionally equivalent or approximant to PC. Similarly, it may be premature to consider evidence of a reduction in neural signal when expectations are met as *diagnostic* of PC.

Acknowledgments

Supported by NSF 1754284, NSF IGERT 1144399, & NSF NRT 1747486 (PI: J.S.M.).

References

Allopenna, P.D., Magnuson, J.S. & Tanenhaus, M.K. (1998) Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language* 38, 419-439.

Altmann, G.T.M. & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge. *J. Memory & Language,* 57, 502-518.

Blank, H., & Davis, M. H. (2016). Prediction errors but not sharpened signals simulate multivoxel fMRI patterns during speech perception. *PLoS Biology*, *14*: e1002577.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117-1121.

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

Frauenfelder, U. H. & Peeters, G. (1998). Simulating the time course of spoken word recognition: an analysis of lexical competition in TRACE. In J. Grainger & A. M. Jacobs (Eds.), *Localist connectionist approaches to human cognition* (pp. 101-146). Mahwah, NJ: Erlbaum.

Gagnepain, P., Henson, R.N., Davis, M.H. (2012) Temporal predictive codes for spoken words in human auditory cortex. *Current Biology, 22*(7), 615-621.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Percep. & Psychophyss, 28,* 267-283.

Hickock, G. (2012). The cortical organization of speech processing: Feedback control and predictive coding the context of a dual-stream model. *Journal of Communication Disorders, 45,* 393-402.

Kuperberg, G.R. & Jaeger, T.F. (2015). What do we mean by prediction in language comprehension? *Language & Cognitive Neuroscience, 31(1)*, 32-59.

Lillicrap, T.P. & Santoro, A. (2019). Backpropogation through time and the brain. *Current Opinion in Neurobiology, 55*, 82-89.

Magnuson, J. S., Mirman, D., Luthra, S., Strauss, T., & Harris, H. (2018). Interaction in spoken word recognition models: Feedback helps. *Frontiers in Psychology*, 9:369.

Magnuson, J.S., Tanenhaus, M.K., & Aslin, R.N. (2008). Immediate effects of form-class constraints on spoken word recognition. *Cognition, 108(3),* 866-873.

McClelland, J. L. (2013). Integrating probabilistic models of perception and interactive neural networks: A historical and tutorial review. *Frontiers in Psychology, 4*, 503.

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

Nieuewland, M.S., Politzer-Ahles, S., Heyselaasr, E., Segaert, K., Darley, E., Kazanina, N. … Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife, 7*: e33468. doi:10.7554/eLife.33468.

Rao, R. & Ballard, D. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2(1),* 79-87.

Sohoglu, E. & Davis, M.H. (2016). Perceptual learning of degraded speech by minimizing prediction error. *Proc. Nat'l Academy Sci., 113(12)*, E1747-56.

Strand, J., Brown, V., Brown, H., & Berg, J. (2017). Keep listening: Grammatical context reduces but does not eliminate activation of unexpected words. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 44,* 962-973. doi: /10.1037/xlm0000488

Yildiz, I.B., von Kriegstein, K., & Kiebel, S.J. (2013). From birdsong to human speech recognition: Bayesian inference on a hierarchy of nonlinear dynamics systems. *PLoS Computational. Biology, 9(9)*: e1003219.