Predictive coding explains dissociable neural and behavioral signatures of sentence processing. Samer Nour Eddine, Sathvik Nair, Alba Jorquera Jimenez De Aberasturi, Jonathan Simon, Philip Resnik; University of Maryland, College Park

Words that are unexpected in sentence context require more cognitive effort to process, as reflected in slower behavioral responses (e.g. reading time) and increases in evoked neural activity (N400 amplitude). Although behavioral and neural measures often pattern together, there are also important differences between the two that suggest that they index different aspects of the comprehension process[1,2]. A fundamental goal in psycholinguistics is to characterize the sentence processing algorithm that underlies these effects, potentially capturing both their similarities but also their differences. No existing process model of language comprehension has successfully accounted for both reading times and N400 responses within a unified framework. Here, we show that predictive coding[3,4]—a biologically plausible algorithm proposed to implement perceptual inference in the brain—provides such a framework.

Method. We utilize a temporal predictive coding model of sentence processing[5,6]. This model processes each word in a sentence by iteratively updating a latent state representation over time to minimize two prediction errors: a *bottom-up* error, i.e. the difference between the word input and its reconstruction from the current latent state; and a *temporal* prediction error, i.e. the difference between the current latent state and its prediction from the previous latent state. The model infers the best explanation for the current word given its context by converging to a stable latent state that minimizes these errors. We use GloVe embeddings as word inputs and trained the model on the Brown corpus. As a proxy for a word's reading time, we use the number of iterations needed for the latent state to converge to a stable value. As a proxy for the N400, we compute the cumulative influence of bottom-up prediction error on latent state updates across inference (see Algorithm 1). Using a dataset of reading times and N400s for identical materials[7], we hypothesized that the analogous model-derived measures would explain unique variance in each. Specifically, we regressed go-past reading times and N400s on model-derived measures.

Results. We find a double dissociation between the model-derived measures and the behavioral and neural outcomes (see Table 1). For reading times, a greater number of iterations to convergence was significantly associated with slower reading times, and likelihood ratio tests confirmed that removing this predictor significantly reduced model fit. By contrast, cumulative bottom-up influence was not predictive of reading times, and its removal did not harm model fit. For N400 amplitudes, the pattern reversed: larger cumulative bottom-up influence was significantly associated with more negative N400 amplitudes, and likelihood ratio tests confirmed its contribution to model fit. Number of iterations to convergence was not associated with N400 amplitudes, and removing it did not reduce model fit.

Discussion. These results demonstrate that predictive coding can account for both neural and behavioral measures of sentence processing. Crucially, both measures derive from a single underlying inference algorithm (see Algorithm 1), yet they capture functionally distinct aspects of that process, potentially explaining why behavioral and neural measures have different properties: the number of iterations to convergence indexes the total time required for the system to integrate bottom-up input with contextual expectations into a stable interpretation, mapping naturally onto go-past reading times; in contrast, the cumulative bottom-up influence indexes how much information the input had to contribute to activate representations not pre-activated by prior context—which the N400 has been argued to reflect[1,4,8]. More broadly, this demonstrates that predictive coding offers a promising, interpretable framework for studying brain-behavior relationships in comprehension, and for generating testable hypotheses about when and why neural and behavioral measures diverge.

Outcome	Model-derived predictor	β range	*p* range	χ² (LRT)	*p* (LRT)
RT	# Iterations to convergence	8.2 to 9.1	**.001–.003**	8.64–10.35	**.001–.003**
RT	Total bottom-up influence	5.9 to 8.3	.07–.19	1.72–3.27	.07–.19
N400	# Iterations to convergence	-0.03 to -0.06	.19–.50	0.46–1.68	.20–.50
N400	Total bottom-up influence	-0.18 to -0.20	**.003–.007**	7.38–8.63	**.003–.007**

Table 1. Mixed-effects regression results testing whether model-derived measures explain unique variance in N400 amplitudes and go-past reading times. All predictors were z-scored. Due to collinearity, each model-derived predictor was tested under two residualization schemes: residualized on control variables only, or residualized on controls plus the other predictor. Ranges reflect estimates across both schemes. β = standardized coefficient; χ² = likelihood ratio test (LRT) comparing the full model to a model with that predictor removed. Control variables included word length, orthographic neighborhood density (OLD20), position in sentence, and cosine distance of GloVe embedding of the current word input from the previous word input. Random intercepts were fit for each word.

Algorithm 1: Predictive Coding Inference Linking Reading Time and N400

for each word w_t:
$y \leftarrow \textsc{Embedding}(w_t)$
// Initialize latent state. If $t = 0$, x *is randomly initialized*
$x \leftarrow x_{t-1}$
// Temporal prediction, i.e. predict where latent state will converge
$\hat{x} \leftarrow W_{xx}x_{t-1} + b_x$
Iters, $\mathcal{N} \leftarrow 0$
while $\|\Delta_x\| >$ threshold:
// Input reconstruction
$\hat{y} = W_{xy}x + b_y$
// Prediction errors
$\varepsilon_{\text{bottom-up}} = y - \hat{y}$
$\varepsilon_{\text{temporal}} = x - \hat{x}$
// Gradient components
bottom_up $= W_{xy}^{\top}\varepsilon_{\text{bottom-up}}$
temporal $= -\,\varepsilon_{\text{temporal}}$
// Latent update
$\Delta_x =$ temporal + bottom_up
$x \leftarrow x + \alpha\Delta_x$ // α is a fixed learning rate
// Cumulative bottom-up influence
$\mathcal{N} \leftarrow \mathcal{N} + (\alpha\Delta_x)^{\top}$bottom_up
Iters ← Iters + 1
end while
ReadingTime(w_t) = Iters
N400$(w_t) = \mathcal{N}$
$x_t \leftarrow x$

Algorithm 1. Word-by-word inference under the trained temporal predictive coding model. For each word, the latent state is iteratively updated to minimize bottom-up and temporal prediction errors. Reading time is defined as the number of iterations to convergence, and the N400 is defined as the cumulative influence of the bottom-up error on the latent state change during inference.

References. [1] Kutas & Federmeier, 2011. *Annual Rev. Psych.*; [2] Rayner & Clifton, 2009. *Biol. Psych.* [3] Rao & Ballard, 1999. *Nature Neuro.* [4] Nour Eddine, Brothers, Wang, Spratling, & Kuperberg, 2024. *Cognition*. [5] Millidge, Tang, Osanlouy, Harper, & Bogacz, 2024. *PLOS Comp. Bio.* [6] Ohams, Nair, Bhattasali, & Resnik, 2026. *JML.* [7] de Varda, Marelli, & Amenta, 2023. *Behav. Res. Methods.* [8] Nour Eddine, Brothers, & Kuperberg, 2022. *Psych. Learning Motiv.*