

## From Feedforward to Feedback: Converging evidence from M/EEG and predictive coding simulations of residual information flow in the language system

**Introduction.** A large body of electrophysiological work has shown that semantically unexpected words elicit an N400 evoked response between 300–500ms [1]. In plausible sentences, this is often followed by a Late Frontal Positivity (LFP) between 600–1000ms [2,3]. It has been hypothesized that the N400 reflects a predominantly *feedforward, bottom-up sweep* in which unpredicted lexico–semantic information is propagated up the left temporal-frontal hierarchy, whereas the LFP reflects a *retroactive, top-down feedback sweep*, in which higher-level updates are propagated down the hierarchy to realign lower-level lexical representations [4]. Here, we took two complementary approaches to test this hypothesis. First, we carried out computational simulations to ask whether the proposed bottom-up and top-down sweeps can be mapped onto distinct elements within an implemented predictive coding model of language processing [5]. Second, we conducted a combined MEG–EEG study to test for a reversal in the net direction of dipole current flow (feedforward to feedback) between 300–500ms and 600–1000ms within the same cortical regions. Such a polarity shift is expected based on cortical laminar microcircuitry, because feedforward inputs to middle layers and feedback inputs to superficial layers drive dendritic currents in opposite directions [6,7] (Figure 1).

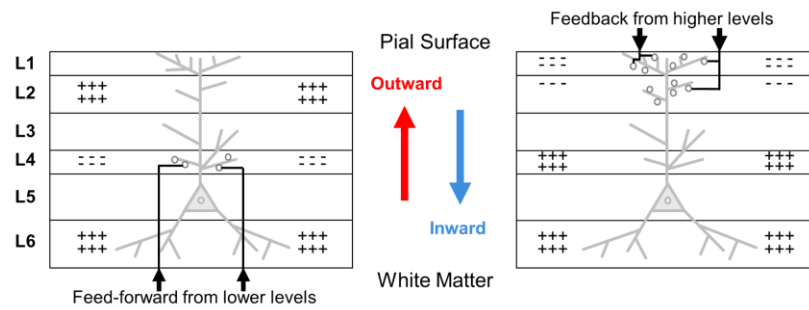
**Experiment 1 (PC Simulations):** We implemented a two-unit hierarchical PC model of lexico-semantic processing using an optimization algorithm that approximates Bayesian inference [5] (Figure 2). In the Expected condition, we provided top-down pre-activation (90% pre-activation of target) to the highest conceptual layer for 20 iterations, then provided the target orthographic input to the lowest layer for another 20 iterations. In the Unexpected condition, we used the same orthographic input, but preceded it with 20 iterations in which the highest layer carried a uniform prior (<1% pre-activation of target). We operationalized the N400 as *bottom-up lexico-semantic prediction error* (residual input information not captured by prior top-down predictions) and the LFP as *top-down lexico-semantic bias* (residual information present in higher-level states but not yet instantiated in lower-level states). For each quantity, we constructed separate time courses for Expected and Unexpected inputs by plotting its magnitude at each iteration as the algorithm converged. As shown in Figure 2, we found that lexico-semantic prediction error closely tracked the time course of the empirical N400 effect, while the top-down bias signal captured the later LFP effect.

**Experiment 2 (M/EEG):** We recorded M/EEG while 31 native English speakers read 210 high-constraint sentences ending with expected words ( $M = 89.1\%$  cloze) and 210 low-constraint sentences ending with unexpected words ( $M = 0.9\%$  cloze), presented with a 600 ms SOA. ERPs confirmed larger N400s and LFPs to unexpected than expected words (Figure 3). Additionally, we localized MEG responses using polarity-preserving, orientation-constrained source reconstruction methods, and used cluster-based permutation tests between 300–500ms to localize the N400 expectancy effect. This analysis revealed significant clusters (*Unexpected > Expected*) within left superior/mid temporal, inferior frontal, and ventromedial temporal cortices. For each cluster, we then extracted the full source-level time courses and asked whether the same vertices were reactivated (*Unexpected > Expected*) between 600–1000ms with a reversed dipole polarity. This analysis revealed significant reversed dipole effects in all three regions.

**Conclusion.** Together, these predictive coding simulations and MEG dipole reversals provide converging computational and neurophysiological evidence for a reversal of information flow across the language hierarchy between 300–500ms and 600–1000ms. They point to a dynamic interplay in which an initial feedforward sweep carries unexpected lexical information up from left temporal to inferior frontal cortices, while later retroactive feedback sharpens lower-level lexical representations in light of higher-level inferences. In this way, we situate the N400 and LFP effects during language comprehension as complementary markers of recurrent information exchange across the cortical hierarchy [8].

# From Feedforward to Feedback: Converging evidence from M/EEG and predictive coding simulations of residual information flow in the language system

**Figure 1.** Schematic of cortical laminae, showing how feedforward signals (left) arise from deep layers and generate an outward-oriented dipole, and feedback signals (right) terminate in superficial layers, generating an inward-oriented dipole on the scalp surface.



**Figure 2.** Schematic depiction of the hierarchical predictive coding model (left) and the simulated bottom-up and top-down effects for **Unexpected** and **Expected** items (right).

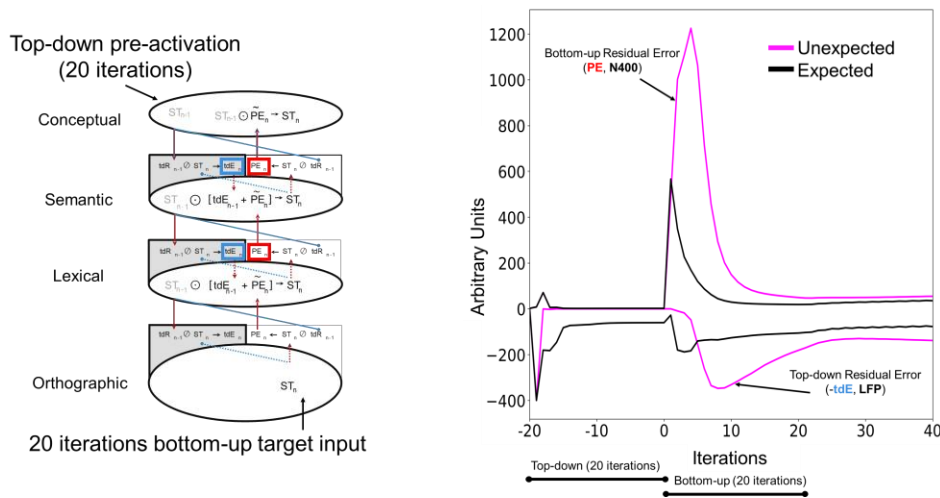
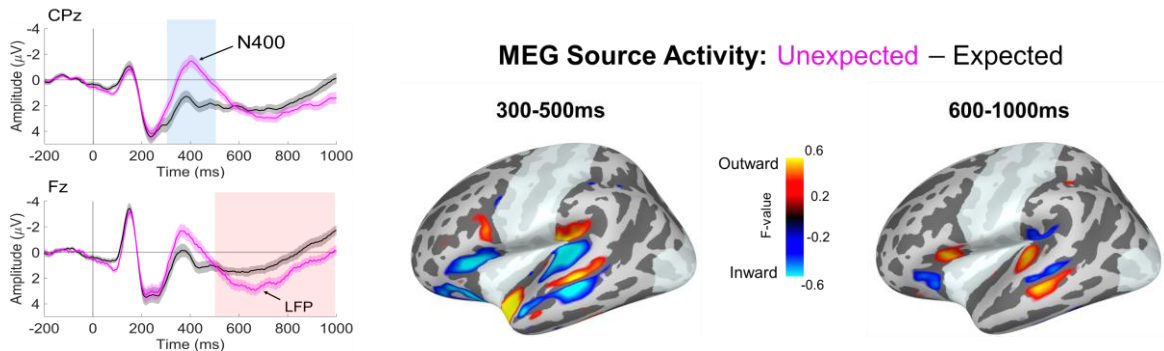


Figure 3. ERPs showing larger N400s and LFPs to **Unexpected** vs **Expected** words (left) and MEG source localized dSPMs showing significant patches of activity in mid/superior temporal, inferior frontal, and ventromedial temporal regions.



**References:** [1] Kutas & Hillyard (1984). *Nature*; [2] Federmeier et al., (2007). *Brain Research*; [3] Kuperberg, Brothers, & Wlotko (2020). *JCN*; [4] Wang et al., (2023). *Cerebral Cortex*; [5] Eddine et al., (2024). *Cognition*; [6] Næss et al., (2021). *NeuroImage*; [7] Bastos et al., (2012). *Neuron*; [5] Lee & Mumford (2003). *JOSA*