What the brain tells us about language comprehension:
A hierarchical generative framework
Gina Kuperberg MD PhD

I will discuss neuroimaging evidence that supports a hierarchical dynamic generative framework of language comprehension (Kuperberg & Jaeger, 2015; Kuperberg, 2016). Within this framework, the comprehender constantly generates *hypotheses* about the underlying message that she believes that the producer is intending to communicate, and generates top-down probabilistic predictions at multiple lower level representations in order to test these hypotheses against new bottom-up evidence provided by the unfolding bottom-up input. These high-level hypotheses can be conceptualized as lying at the top of an internal hierarchically organized *generative model* — the network of linguistic and non-linguistic representations that, at any given time, the comprehender believes can best explain the statistical properties of the bottom-up input that she has thus far encountered, given her beliefs about the broader statistical structure of her current environment as well as her communicative goals. As new bottom-up evidence becomes available, the comprehender learns whether her probabilistic predictions at each level of representation within the generative model are supported. Bottom-up evidence that has not already been predicted at a given representational level constitutes *prediction error*, and is passed up the generative model and used to update her higher-level hypotheses through Bayesian inference. I will suggest that prediction error at different levels of representation manifests in the brain as distinct spatiotemporal neural signatures. Specifically, neural activity within the left anterior temporal cortex, observed between 300-500ms after the onset of unpredicted inputs, and corresponding to the N400 ERP effect, may reflect prediction error at the level of semantic features, while later activity within the left inferior frontal cortex, observed between 400-700ms and corresponding to a later anteriorly distributed negativity ERP effect, may index prediction error at the level of event structures (representations of 'who does what to whom'). Incremental modulation of this temporal-frontal neural network may therefore reflect iterative cycles of probabilistic prediction and inference that proceeds until prediction error across the entire generative model is minimized and the comprehender has converged upon the particular message-level representation that best explains the bottom-up input.

Importantly, I will argue that these generative models are not fixed, and that comprehenders can modify their structure, or switch to (or infer) alternative, previously stored models, in rapid response to changes in the statistical structure of their broader communicative environment and/or their communicative goals. I will suggest that such adaptation manifests in the brain as an additional set of spatiotemporal neural signatures that are distinguished by the levels of representation at which adaptation takes place. Specifically, a late frontally-distributed positive ERP waveform may reflect adaptation at the interface between semantic and word-form representations, while a late posteriorly-distributed positive ERP waveform (the P600) may reflect adaptation at the interface between semantic and event structure representations. I will conclude by discussing the implications of this dynamic generative architecture for linking comprehension, production and learning in healthy individuals, and how these links might break down in neuropsychiatric disorders such as schizophrenia (Brown & Kuperberg, 2015).

Kuperberg GR, Jaeger TF. Language, Cognition & Neuroscience. 2015
Kuperberg GR. Language, Cognition & Neuroscience. 2016.
Brown M. Kuperberg GR. Frontiers in Human Neuroscience. 2015.