Generative Prior in Scalar Inference via Free-Energy Principle

Standard accounts of adjectival vagueness, such as the Rational Speech Act (RSA) framework, model meaning composition as a recursive Bayesian inference problem. In this model, the pragmatic listener L_1 infers the world state from an utterance by simulating a speaker S_1 and a literal listener L_0 (Lassiter & Goodman, 2015). While successful in many contexts, RSA faces challenges with scalar implicature, where listeners often appear insensitive to prior manipulations. To remedy this, the "wonky world" extension (wRSA) introduces a latent variable w that modulates the state prior $P(s|w)$, allowing listeners to back off to a uniform distribution when an utterance is odd (Equations 1–3; Degen et al., 2015). However, the dynamics of this switch remain under-specified, as $P(w)$ is typically stipulated rather than derived from cognitive constraints.

$$P_{L_0}(s \mid u, w) \propto [\![u]\!](s) \cdot P(s \mid w) \tag{1}$$

$$P_{S_1}(u \mid s, w) \propto \exp(\lambda \ln P_{L_0}(s \mid u, w)) \tag{2}$$

$$P_{L_1}(s, w \mid u) \propto P_{S_1}(u \mid s, w) \cdot P(s \mid w) \cdot P(w) \tag{3}$$

This project aims to develop a neurobiologically grounded (toy) model that maps RSA's recursive architecture to the Free Energy Principle (FEP) and the Predictive Coding (PC) framework. The FEP posits that biological systems minimize variational free energy by continuously updating internal models to suppress surprisal (Friston, 2010). PC enacts this mechanism by employing hierarchical message passing, where top-down predictions attempt to 'explain away' sensory inputs while bottom-up prediction errors drive updates to the internal model. Experimentally, low-frequency oscillations are correlated with top-down prediction maintenance and high-frequency oscillations with bottom-up integration of prediction errors. To map RSA onto this framework, I propose that S_1 corresponds to the brain's top-down process, which predicts utterances from hidden world states, while L_1 corresponds to the bottom-up process of model inversion. Note that S_1 and L_1 here are no longer individual levels, but rather the bi-directional connections between the top level and bottom level. This mapping is supported by recent evidence that active world-model construction in speech processing is better indexed by non-phase-locked oscillatory dynamics (Wu et al., 2025).

My hypothesis is that low informativity is registered as surprisal at top level and in turn down-regulates the world-model of the level below. Scalar implicature's apparent insensitivity to priors is an emergent consequence of precision weighting across this hierarchy. In a scenario with an extremely skewed prior (e.g., "All marbles sink" is certain), the utterance "(at least) Some marbles sank" yields negligible utility (Equation 4), which is quantified by cost subtracted from the Kullback-Leibler divergence between the posterior and prior (Goodman & Frank, 2016). In the proposed model, **the prediction error generated by this low utility dynamically down-regulates the precision (inverse variance, Σ^{-1}) of the current world prior.** This process is defined by the gradient ascent of Free Energy (Equation 5; Bogacz, 2017):

$$\text{Utility} = D_{KL}(P||Q) - \text{Cost} = \sum_x P(x) \ln \frac{P(x)}{Q(x)} - \text{Cost} \tag{4}$$

$$\Delta\Sigma_p \propto \frac{\partial F}{\partial \Sigma_p} = \frac{1}{2}(\epsilon_p \epsilon_p^T - \Sigma_p^{-1}) \tag{5}$$

Thus, the "wonky" switch (w) emerges naturally from free-energy minimization, potentially observable experimentally as beta-band oscillations marking the revision of the world model.

New data on absolute adjectives provide a complementary test case for how the model handles informativity under varying prior constraints. Following Xiang et al. (2022), I contrast absolute adjectives (e.g., *straight*) with relative ones (e.g., *tall*). Xiang et al.'s experimental data show that absolute adjectives are encoded with low-variance world-model priors (high precision). **I hypothesize that this high precision inherently imposes a higher threshold of informativity**: for an utterance to yield significant KL-divergence (Equation 4) against a sharp prior, it must be highly specific, necessitating a more categorical posterior distribution. This creates a principled symmetry with scalar implicature: whereas scalar inference requires increasing prior variance to make sense of under-informative utterances, absolute adjectives strictly maintain low variance, driving the system toward endpoint-oriented interpretations. By simulating these dynamics, this project aims to unify different types of scalar inference within a single mechanism, linking formal pragmatics directly to neurocomputational process models.

References

Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *Journal of Mathematical Psychology, 76*, 198–211.

Degen, J., Tessler, M. H., & Goodman, N. D. (2015). Wonky worlds: Listeners revise world knowledge when utterances are odd. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.

Friston, K. (2010). The free-energy principle: a unified brain theory?. *Nature Reviews Neuroscience, 11*(2), 127-138.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences, 20*(11), 818–829.

Lassiter, D., & Goodman, N. D. (2015). Adjectival vagueness in a Bayesian model of interpretation. *Synthese, 194*(10), 3801–3836.

Wu, H., Rao, X., & Cai, Z. G. (2025). Probabilistic adaptation of language comprehension for individual speakers: evidence from neural oscillations. *Social Cognitive and Affective Neuroscience, 20*(1), nsaf085.

Xiang, M., Kennedy, C., Xu, W., & Leffel, T. (2022). Pragmatic reasoning and semantic convention: A case study on gradable adjectives. *Semantics and Pragmatics*.