Probable World Semantics as a Solution to Linguistic Ambiguity

Language comprehension requires resolving pervasive ambiguity: pronouns can refer to countless entities, words have multiple meanings, and sentences admit thousands of syntactic parses. Yet humans navigate this ambiguity effortlessly. We propose that comprehenders resolve linguistic ambiguity using Probable World Semantics (PWS) - a probabilistic extension of possible world semantics (Kripke, 1980; Montague, 1980) where interpretations are chosen in proportion to their a priori probability of being true in the world. We formalize PWS as:

$$P(\text{meaning}_i|\text{utterance}) \propto P(\text{utterance}|\text{meaning}_i) \times P(\text{meaning}_i|\text{world})$$

where $P(\text{meaning}_i|\text{world})$ reflects the comprehender's intuitive probability that meaning_i might actually be true in the world. We PWS primarily pronoun interpretation, including both pronoun biases (Garvey & Caramazza, 1974; Kehler et al., 2008) and Winograd Schema (Levesque et al., 2012), where changing a single word flips pronoun reference. We compare to computational versions of classic pronoun interpretation theories such as Centering Theory (Grosz et al., 1995), Causal Closeness (Sagi & Rips, 2014), and two variants of a “Many Factors” model that captures key aspects of recent psycholinguistic accounts (Arnold, 2025; Hartshorne et al., 2015; Kehler & Rohde, 2019).[1] In order to provide a high benchmark against which to compare PWS, we also test three LLMs known to perform well on Winograd schema (T5, OLMo-2-1124-13B-Instruct, Gemma 3 27B).

Experiment 1 compared PWS predictions to human pronoun judgments across 240 Winograd sentence pairs and 40 pronoun bias stimuli. PWS achieved high performance (r = .70 for Winograd sentences, r = .71 for pronoun biases), substantially outperforming the linguistic models and matching or exceeding three state-of-the-art Large Language Models (T5, OLMo, Gemma). Full results for pronoun biases are shown in Fig. 1 (space precludes full results for the other evaluations.)

While the PWS model in Experiment 1 relied on human judgments of $P(\text{meaning}_i|\text{world})$, **Experiment 2** focused on two scenarios (tug-of-war games and marble collisions) for which we have highly accurate models of human beliefs about real-world probabilities. Using these probabilities, PWS achieved exceptional accuracy (r = .91). All other models were at chance, except Gemma (r = .35).

Experiments 3a-3b provided direct causal evidence by manipulating participants' world knowledge. When told that students from certain schools are stronger (Exp. 3a) or that weak players win at tug-of-war due to divine intervention (Exp. 3b), pronoun interpretation shifted predictably according to PWS (r = .68 and r = .97, respectively) while other models and LLMs largely failed.

Experiment 4 extended PWS to scenarios where event participants have differential likelihoods of engaging in described events (e.g., drunk vs. sober drivers causing accidents). PWS successfully integrated both event-type probabilities and participant-specific priors (r = .76), significantly exceeding performance by all models except T5 (r = .71).

Experiments 5-6 demonstrated PWS's broader applicability beyond pronouns. Experiment 5 showed that PWS could predict pronoun interpretation even without explicit discourse connectives by marginalizing over possible discourse relations (r = .93) – probabilities of discourse relations were themselves calculated through PWS – significantly outperforming all other models. **Experiment 6** extended PWS to distinguish causal from epistemic explanations in ambiguous "because" constructions (r = .95). Other models made no predictions.

Across six experiments involving over 1,400 stimuli, PWS consistently provided the most accurate account of human linguistic behavior, often achieving near-ceiling performance while competing models failed catastrophically on multiple evaluations. Crucially, PWS makes novel, testable predictions that were confirmed experimentally (e.g., Exps. 5-6).

These results suggest that language comprehension fundamentally depends on probabilistic reasoning about possible worlds. Rather than relying solely on syntactic or semantic rules, comprehenders integrate linguistic input with their beliefs about what is likely to be true, selecting interpretations in proportion to the *a priori* probability that the interpretation would be true of the world. PWS thus provides a unified computational account of diverse phenomena in language understanding, from pronoun resolution to discourse structure inference, while revealing the deep connection between sentence interpretation and world knowledge in human cognition.

---

[1] While much work has focused on the differences between these recent psycholinguistic accounts, for present purposes those differences are sufficiently minor and can be elided.

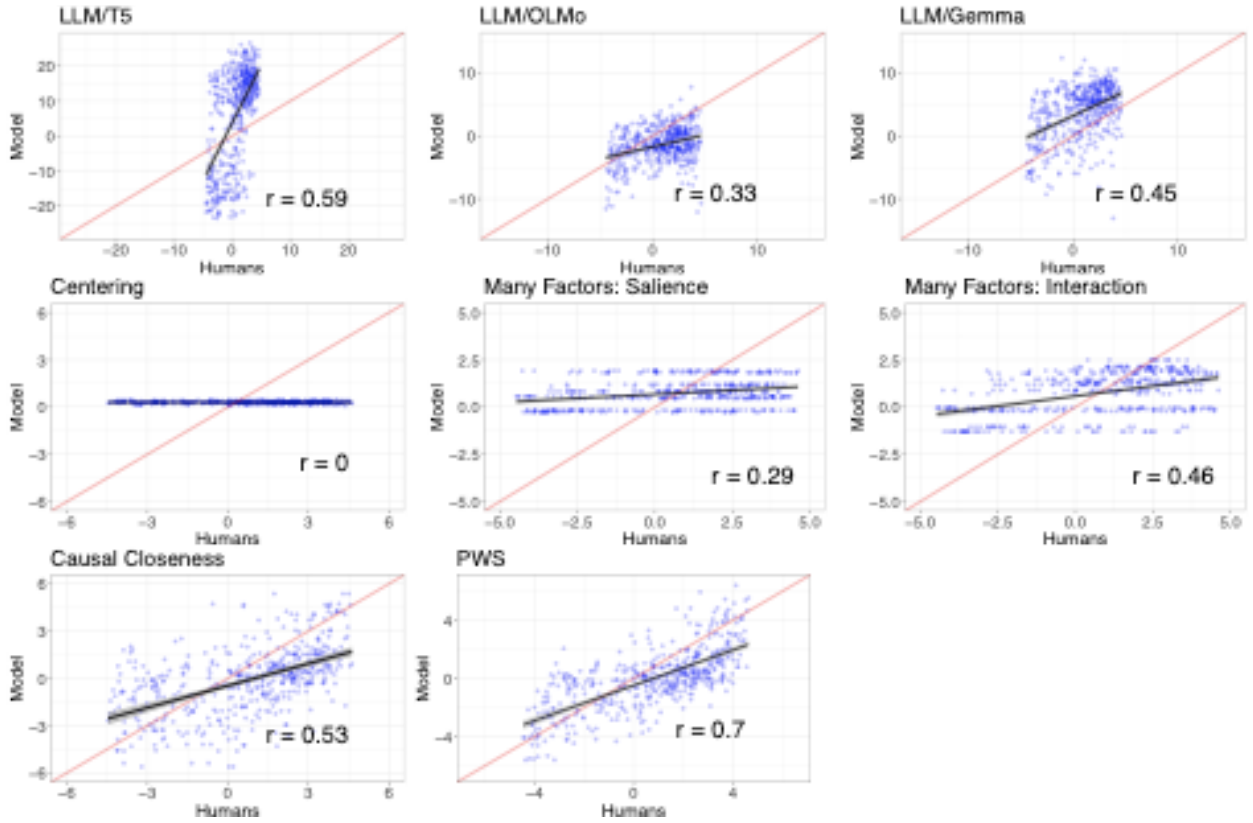

Fig 1. Exp. 1 results for pronoun biases: log-odds pronoun refers to subject vs. object for the three LLMs (top row), four psycholinguistic models, and PWS. Each point represents a stimulus. The black line indicates the best linear fit between model and data, with the 95% confidence interval indicated in gray. The red diagonal line indicates what a perfect fit between model and data would look like.

Arnold, J. E. (2025). Why does recency guide pronoun comprehension? It's not just topicality, attention, or predictability. Discourse Processes, 1-23.

Garvey, C., & Caramazza, A. (1974). Implicit causality in verbs. *Linguistic inquiry*, 5(3), 459-464.

Grosz, B. J., Joshi, A., & Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. Computational linguistics, 21(2), 203-225.

Hartshorne, J. K., O'Donnell, T. J., & Tenenbaum, J. B. (2015). The causes and consequences explicit in verbs. Language, cognition and neuroscience, 30(6), 716-734.

Kehler, A., Kertz, L., Rohde, H., & Elman, J. L. (2008). Coherence and coreference revisited. Journal of semantics, 25(1), 1-44.

Kripke, S. A. (1980). *Naming and necessity*.

Levesque, H., Davis, E., and Morgenstern, L. (2012). The Winograd schema challenge. In Proceedings of the 13th International Conference on the Principles of Knowledge Representation and Reasoning.

Kehler, A., & Rohde, H. (2019). Prominence and coherence in a Bayesian theory of pronoun interpretation. Journal of Pragmatics, 154, 63-78.

Montague, R. (1980). *Universal grammar*.

Sagi, E., & Rips, L. J. (2014). Identity, causality, and pronoun ambiguity. Topics in Cognitive Science, 6(4), 663-680.