Modeling human inferences and reading behavior with an incremental, resource-rational model of noisy-channel language processing

Thomas Hikaru Clark, Jacob Hoover Vigly, Edward Gibson, Roger Levy
Department of Brain and Cognitive Sciences, MIT

Background. How are comprehenders able to extract meaning from atypical utterances, such as those produced by people with aphasia (PWAs) or L2 speakers, but also on occasion by fluent speakers? Noisy Channel theory [1,2] provides a Bayesian inference account – comprehenders may interpret utterances non-literally in favor of an alternative with higher prior probability. However, we lack implemented computational models of prior expectation and error likelihood capable of predicting human processing (e.g., inferences and reading times) of arbitrary utterances. Here, we use particle filters [3] to model noisy-channel inference subject to constraints on cognitive resources [4]. We test whether the following algorithmic predictors provide an improved fit to human reading behavior for ill-formed utterances: 1) surprisal calculated under noisy-channel inference, and 2) the probability of finding alternative interpretations (quantified using our model). **Model.** We use a particle filter to model sentence processing for "noisy" utterances as incremental probabilistic inference over intended strings and production errors. The particle filter makes use of a *language model* (LM) defining a prior over intended strings (here a simple n-gram model trained on typical, "clean" utterances), and an *error model* defining possible errors and their probabilities (e.g. substitution, deletion, backtracking). Each particle corresponds to a hypothesis about the identity of the intended sentence and errors; particle weights are updated in light of each new observation (word). Diversity is added to particles via *rejuvenation* steps which propose alternative values for intended sentences and errors, and which are probabilistically accepted or rejected based on prior probability and error likelihood using the Metropolis-Hastings algorithm. We also derive a novel model-based predictor, rejuvenation acceptance rate (RAR), as the proportion of rejuvenations which were probabilistically accepted by the particle filter at each time step, which quantifies the ease of proposing new explanations for a given observation; we test RAR as a proxy for incremental integration cost, alongside surprisal (*Fig. 1*). In this study all utterances are grounded in the AphasiaBank [5] "Window" picture-description task, in English. We validate the model on a subset of AphasiaBank with 100 randomly sampled utterances each from PWAs and control speakers. For both groups, our model had a lower per-word surprisal than the n-gram baseline (b=-0.963, 95% CI=[-1.320, -0.606], p<.001), with a larger difference for PWA utterances (b=0.754, 95% CI = [0.284, 1.224], p=.002) (*Fig. 2*), indicating that our model improves the modeling of out-of-distribution language in general, and impaired language especially, even without explicit training on impaired utterances. **Behavioral Experiment.** We use the Mouse Tracking for Reading [6] paradigm to collect per-word reading time (RT), skipping, and regression measures from N=50 native English-speaking participants on Prolific, on a total of 240 utterances with a variety of introduced errors. Participants also saw an illustration of the "Window" scene for context, and responded with their inference about the intended sentence after reading. We use linear mixed models to predict RTs and logistic mixed models to predict skips and regressions. Each model had one of the following target predictors: "noisy-channel" surprisal from the particle filter, n-gram surprisal, GPT-2 surprisal (a known strong predictor of RTs [7]), and two varieties of RAR (differing in whether they propose alternatives using the literal utterance or the current inferred intended sentence), alongside word length and frequency, with by-word and by-speaker random intercepts. We compare predictive power with per-observation change in Log-Likelihood (ΔLogLik) (*Fig. 3*). **Results.** Noisy-channel surprisal was more predictive of every continuous RT measure and first-pass skips than n-gram surprisal, but fell short of GPT-2 surprisal's predictive power. However, noisy-channel and n-gram surprisal both outperformed GPT-2 at predicting first-pass regressions. Additionally, for every response variable, at least one of the RAR predictors had higher ΔLogLik than noisy-channel surprisal and offered the overall best ΔLogLik for first-pass skips, non-first fixation times, and regressions. **Conclusion.** Our results provide evidence that explicitly modeling rational inference over possible errors improves the fit of a next-word-prediction model to behavioral data. Our model provides a step towards an interpretable, algorithmic account of language processing that is robust to potential errors.

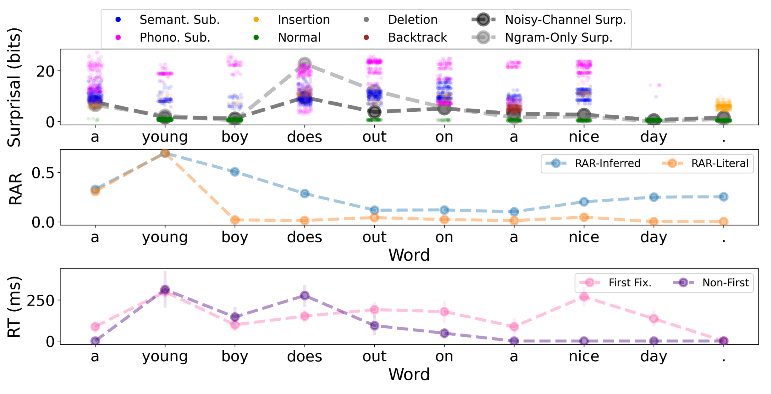

Fig. 1: **Model-based predictors and human RTs.** Comparison of surprisal from baseline n-gram LM (**gray**) and noisy-channel particle filter model (**black**), two varieties of Rejuvenation Acceptance Rate (RAR), and human RT measures. In the surprisal panel, colored points denote individual particle surprisals; colors denote inferred error type. In this example, the noisy-channel model fails to predict the word "*does*" given the context (no **normal** actions among the particles), while the lowest-surprisal particles correspond to a **phonological substitution**. For results comparing the predictive performance of model-based predictors for human reading measures, see Figure 3.

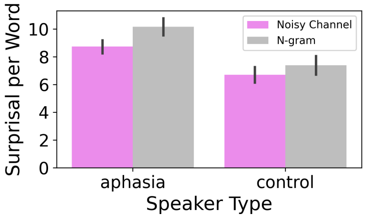

Fig. 2: **Model validation.** Per-word surprisal comparison for 100 randomly sampled PWA and control speaker utterances from the AphasiaBank *Window* picture-description task according to baseline n-gram LM (**gray**) trained on disjoint control speaker utterances, and our noisy-channel model (**magenta**). There is both a main effect of model type and an interaction between model type and speaker type.

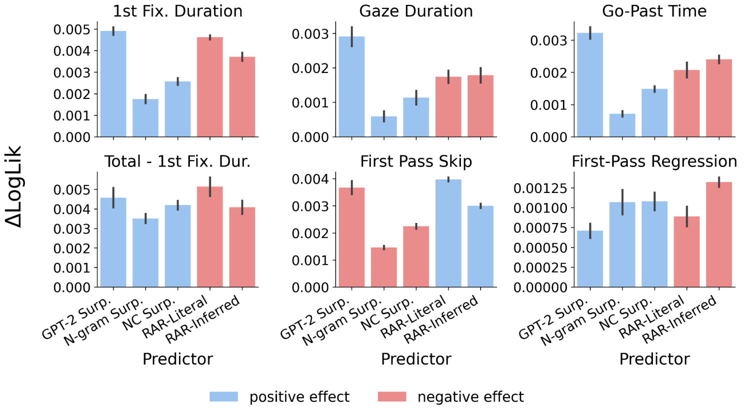

Fig. 3: **Predicting reading behavior.** Per-observation ΔLog-Likelihood comparison (higher is better) of model-based predictors across 4 RT measures, first-pass skips, and first-pass regressions. Error bars denote 95% CIs of the ΔLog-Likelihood across 5 folds of data. Blue and red bars denote positive and negative effects, respectively. Noisy-channel surprisal ("NC Surp.") provides better fit to humans than n-gram surprisal for all RT measures, and RAR provides even better fit, exceeding that of GPT-2 surprisal for several response variables.

Example Stimulus	Introduced Error	Example Stimulus	Introduced Error
a young boy is out on a nice day .	None	a young boy does out on a nice day .	Phonological Substitution
he is at the front it door .	Insertion	him goes over to the window .	Semantic Substitution
and hes kicking ball .	Deletion	and his dad and his dad was upset .	Backtrack

References: [1] Levy (2008). A Noisy-Channel Model of Human Sentence Comprehension under Uncertain Input. Proceedings of EMNLP. **[2]** Gibson, Bergen, & Piantadosi, (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. PNAS. **[3]** Levy, Reali, & Griffiths (2008). Modeling the effects of memory on human online sentence processing with particle filters. **[4]** Hoover, Sonderegger, Piantadosi, & O'Donnell (2023). The Plausibility of Sampling as an Algorithmic Theory of Sentence Processing. Open Mind. **[5]** MacWhinney, Fromm, Forbes, & Holland (2011). AphasiaBank: Methods for Studying Discourse. Aphasiology. **[6]** Wilcox, Ding, Sachan, & Jäger (2024). Mouse Tracking for Reading (MoTR): A new naturalistic incremental processing measurement tool. JML. **[7]** Shain, Meister, Pimentel, Cotterell, & Levy (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. PNAS