Unexpected events and explicit reference: A noisy-channel RSA model

Ivan Rygaev[1], 1, Martin V. Butz[1], Asya Achimova[1]

[1]University of Tübingen

ivan.rygaev@uni-tuebingen.de

[1] examines how event surprisal affects referring expression production and comprehension. In a production experiment, participants described cartoons in which two colored monsters interacted (e.g., attacked, threw rocks, jumped over, or waved at each other) and one ultimately fell. They produced utterances of the form:

(1) *The yellow monster attacked the red monster and [∅ | it | the yellow/red monster] fell down.*

Speakers used more overt references for surprising outcomes than for expected ones. Referring expressions were influenced by action-based priors (e.g., whether the agent or patient of an attack is more likely to fall), color-based monster strength priors learned during training (the weaker monster should fall), and antecedent role (agent vs. patient). Speakers also produced fewer zero anaphors for surprising events despite their lack of ambiguity, possibly anticipating that listeners in noisy communication settings might infer an ambiguous pronoun instead, yielding a more plausible interpretation (Figure 1, right). The findings were further corroborated in the perception experiment, where participants had to select one of the two outcomes after hearing a noisy audio prompt, either with a pronoun or a zero anaphor (Figure 2, right), and type what they heard.

In this paper, we present a noisy-channel RSA model [2, 3] covering the results of those experiments. Here are the formulas for the literal listener and the pragmatic speaker:

$$P_{L_0}(s \,|\, u_p, a, e) \propto P(s|a) \cdot P(s|e) \cdot \sum_i P(u_p|u_i) \cdot [\![u_i]\!](s) \tag{1}$$

$$P_{S_1}(u_i \,|\, s, a, e) \propto exp(\alpha \cdot (\sum_p P(u_p|u_i) \cdot logP_{L_0}(s|u_p, a, e) - C(u_i))) \tag{2}$$

We consider two states s: $agent\ falls$ and $patient\ falls$, and four utterances: $zero$ (zero anaphor), pro (ambiguous pronoun), $agent$, and $patient$ (noun phrases). Only pro is compatible with both states; the others are unambiguous. Given a perceived utterance u_p, listeners infer the intended utterance u_i using a confusion matrix $P(u_p|u_i)$. We model two error rates $P(pro_p|zero_i)$ and $P(zero_p|pro_i)$, assuming noun phrases are never misheard. The model also includes action priors $P(s|a)$ and monster strength priors $P(s|e)$ as parameters.

We performed a Bayesian data analysis in WebPPL [4] using MCMC [5]. Sampling from uninformative priors and conditioning on observed production and interpretation rates yielded posterior parameter distributions; we report posterior means. The rationality parameter was $\alpha = 3.1$. Estimated inverse error rates were $P(pro_i|zero_p) = 0.6$, $P(zero_i|pro_p) = 0.06$, closely matching the rates in the typing task in the perception experiment despite the model not being trained on those data. Estimated action priors $P(patient\ falls|a)$ were: $jump\ over = 0.43$, $wave = 0.53$, $attack = 0.59$, and $throw\ rock = 0.72$, preserving the ordering from the norming study. The strength prior was $P(s|weaker\ monster\ falls) = 0.59$. Including additional datasets, the model had 17 parameters and was fit to 88 data points[1]. Posterior predictive checks indicated good fit (see Figures 1 and 2), and model comparisons using AIC and BIC supported the inclusion of all parameters.

Surprising events lead speakers to use more explicit references because they anticipate listeners' expectations and perceptual noise. A noisy-channel RSA model quantitatively captures this behavior and recovers independently measured priors. The results demonstrate the value of probabilistic models such as RSA for explaining pragmatic phenomena and linking experimental findings with formal theories of language use.

[1]A shiny app with the full model is available here: https://tinyurl.com/5ex9f9e5

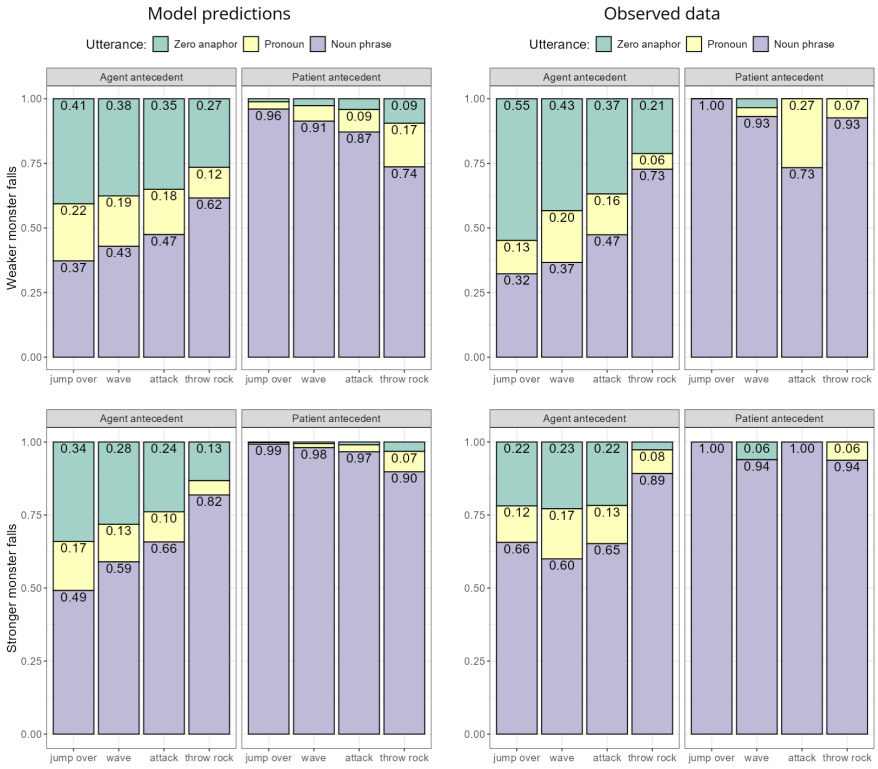


Figure 1: Produced referring expressions. Model predictions (left) and observed rates (right).

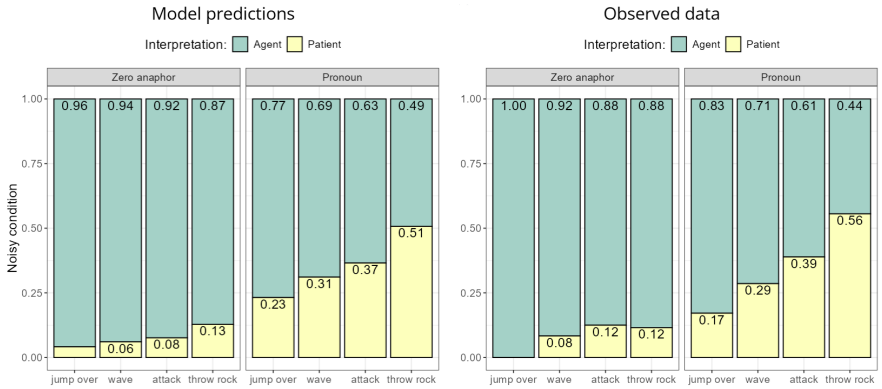


Figure 2: Referring expression interpretation. Model predictions (left) and observed rates (right).

References

[1] Rygaev, I., Demberg, V., Butz, M. V., & Achimova, A. (2026). Event predictability modulates production and comprehension of referring expressions [[submitted for publication]].

[2] Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why bayes' rule is probably important for pragmatics. *Zeitschrift für sprachwissenschaft*, *35*(1), 3–44.

[3] Bergen, L., & Goodman, N. D. (2015). The strategic use of noise in pragmatic reasoning. *Topics in cognitive science*, *7*(2), 336–350.

[4] Goodman, N. D., & Stuhlmüller, A. (2014). The Design and Implementation of Probabilistic Programming Languages [Accessed: 2025-6-24].

[5] Gilks, W. R., Richardson, S., & Spiegelhalter, D. (1995). *Markov chain monte carlo in practice*. CRC press.