RSA model of referential expression production and comprehension under noise

Ivan Rygaev, Martin V. Butz, Asya Achimova
University of Tübingen
ivan.rygaev@uni-tuebingen.de

Imagine a scene where two colored monsters interact (say, one attacks the other). As a result, either the attacker or the attacked monster falls. How can we describe the scene? One way to describe it is in (1):

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

Which expression from the brackets would better fit which outcome, assuming that the fall of one monster is more plausible than the other? Speakers tend to use more overt expressions (e.g. noun phrases) when describing surprising events. They anticipate that listeners would treat ambiguous pronouns according to their prior expectations, and when the outcome contradicts the expectations, speakers try to avoid that [1]. They even avoid conjoined verb phrases ("attacked and ∅ fell down"), even though those are not ambiguous, because supposedly they anticipate that in an imperfect communication channel, the listener might reconstruct a pronoun if it allows arriving at a more plausible interpretation. In a series of experiments (priors assessment for different actions, speech perception under noise, and speech production) we confirmed the intuition outlined above.

Now we present an RSA model [2] partially covering the data. As the results show (Figure 1), when referring to the patient, speakers predominantly use NPs. This aligns with other studies [3]. But then we can expect that listeners would treat other forms, like pronouns, almost exclusively as referring to the agent. However, in a third of cases, a pronoun was taken to refer to the patient. And after excluding the cases of pronoun misperception as zero anaphors under noise, the rate rises to the striking 68%. Can an RSA model account for that?

In our noisy-channel RSA model [4], we assume two possible states of affairs: *agent falls* and *patient falls*, and four possible utterances: *zero* (a zero anaphor, i.e. a conjoined verb phrase), *pro* (an ambiguous pronoun), *agent*, and *patient*. The last two stand for any noun phrases unambiguously referring to the agent or patient respectively. Here is our definition of literal listener L_0, pragmatic speaker S_1, and pragmatic listener L_1:

$$P_{L_0}(s|u) \propto P(s) \sum_i P_{u_i,u} [\![u_i]\!](s) \quad P_{S_1}(u|s) \propto exp(\alpha \sum_i P_{u,u_i} log P_{L_0}(s|u_i)) \quad P_{L_1}(s|u) \propto P(s) \sum_i P_{u_i,u} P_{S_1}(u_i|s)$$

$P(s)$ is a prior probability of s from the priors experiment: $P(agentfalls) = 0.32$ and $P(patientfalls) = 0.68$.

P_{u_i,u_p} is a probability to perceive an intended utterance u_i as utterance u_p. We assume that in noisy conditions, listeners can confuse zero anaphors with pronouns and vice versa at a fixed error rate e, so $P(pro, zero) = P(zero, pro) = e$ and $P(zero, zero) = P(pro, pro) = 1 - e$. This leads speakers to use fewer zero anaphors. Assuming that noun phrases cannot be misinterpreted, we set $P(agent, agent) = P(patient, patient) = 1$.

$[\![u_i]\!](s)$ is the interpretation function, a probability that utterance u refers to state s. For unambiguous utterances, it is categorical: $[\![zero]\!](agentfalls) = [\![agent]\!](agentfalls) = 1$ and $[\![patient]\!](patientfalls) = 1$. For ambiguous pronouns, it is regulated by parameter ab (agent bias): $[\![pro]\!](agentfalls) = ab$ and $[\![pro]\!](patientfalls) = 1 - ab$.

By fitting a computational model in WebPPL [5] we estimated the following optimal parameter values: $e = 0.22, ab = 0.66, \alpha = 2.9$. Figure 2 shows that the posterior predictions from the model fit the data rather well. So, the model was able to cope with the apparent contradiction we described above.

The key to the explanation is the observation that when referring to the agent, speakers also use pronouns rather rarely. Because of that, listeners mostly rely on their priors to disambiguate pronouns. But our experimentally obtained priors are highly biased towards the patient (68% vs 32%). That explains the listeners' behavior.

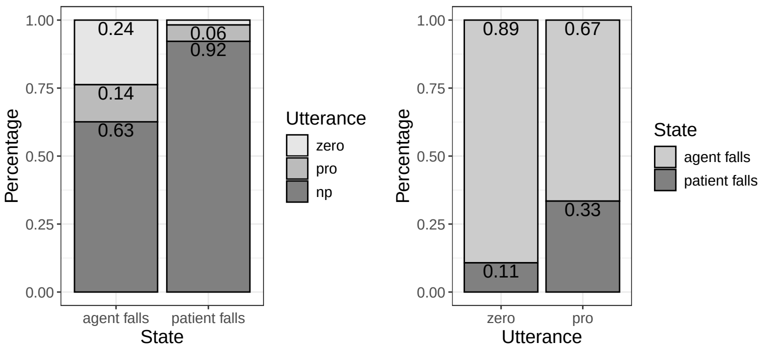

Figure 1: Experimental results. The figure on the left shows the rate of utterances produced for each state. The figure on the rights shows the rate of states selected for each utterance.

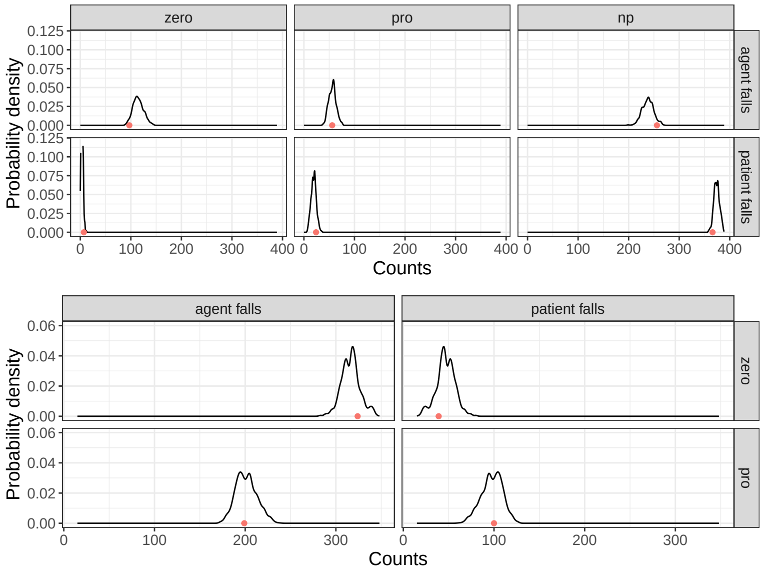

Figure 2: Modeling results (production at the top, perception at the bottom). Red dots show the counts of each condition from the experiment. The black curves show probability distribution of the counts as per the posterior predictions of the model. The red dots lie within the plausible regions of the model predictions.

References

[1] Achimova, A., van Os, M., Demberg, V., & Butz, M. V. (2024). Interpreting implausible event descriptions under noise. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*, 3399–3406.

[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] Demberg, V., Kravtchenko, E., & Loy, J. E. (2023). A systematic evaluation of factors affecting referring expression choice in passage completion tasks. *Journal of Memory and Language*, *130*, 104413.

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

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