Ivan Rygaev

RSA modeling of event predictability and noise in referential expression comprehension

Bandtitel, Hg. v. Nachname, Nachname und Nachname, 20XX, S. XX–YYY
https://doi.orgXXXXXXXXXXX

© 2025 bei Autor:in Ivan Rygaev
Dieses Werk ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz, ausgenommen von dieser Lizenz sind Abbildungen, Screenshots und Logos.

Ivan Rygaev, Universität Tübingen, ivan.rygaev@uni-tuebingen.de, 0000-0003-4135-6954

Zusammenfassung

Experimentelle Daten zeigen, dass Hörer:innnen bei Rauschen im Kommunikationskanal nicht nur bei der Interpretation ambiger Pronomen, sondern auch bei der Interpretation formal eindeutiger Nullanaphern in koordinierten Verbphrasen auf ihre Erwartungen zurückgreifen. Dementsprechend neigen Sprecher:innnen dazu, beim Beschreiben unerwarteter Ereignisse häufiger explizite Ausdrücke (Nominalphrasen) und seltener Nullanaphern zu verwenden. In diesem Artikel stellen wir ein RSA-Modell zum Verständnis referenzieller Ausdrücke vor, das diese Ergebnisse erklärt und sowohl die Vorhersagbarkeit von Ereignissen als auch die Möglichkeit von Rauschen im Kommunikationskanal berücksichtigt.

Schlagwörter: referenzielle Ausdrücke, Vorhersagbarkeit von Ereignissen, Rational Speech Act, Modellierung

Abstract

Experimental data indicate that under noisy conditions, listeners rely on their prior expectations not only when interpreting ambiguous pronouns but also when interpreting formally unambiguous zero anaphors in conjoined verb phrases. Correspondingly, speakers tend to use more explicit expressions (noun phrases) and fewer zero anaphors when describing surprising events. In this paper, we introduce an RSA model of referential expression comprehension that accounts for these findings, incorporating both event predictability and the possibility of noise in the communication channel.

Keywords: referential expressions, event predictability, Rational Speech Act. modeling

1 Introduction

Imagine a scene where two monsters of different colors interact—for instance, one monster throws a rock at the other. As a result, either the attacker or the attacked monster falls. How might we describe such a scene? Some possible descriptions are provided in (1):

(1) The yellow monster threw a rock at the red monster and...

a. ... Ø fell down

b. ... it fell down

c. ... the yellow monster fell down

d. ... the red monster fell down

Which option corresponds better to which outcome, given that one monster's fall is more plausible than the other's?

Listeners generally resolve ambiguous expressions—such as the pronoun *it* in this example—in line with their prior expectations (Winograd, 1972). When hearing (1b), a listener is likely to conclude that the red monster fell, since this outcome aligns with a natural and expected progression of events. Such reliance on prior expectations is central to Bayesian probabilistic reasoning. Bayesian pragmatic frameworks like the Rational Speech Act (RSA) model provide a way to combine prior expectations with linguistic input. When the linguistic signal is clear and compatible with only one world state, the utterance itself drives interpretation. But when the utterance is ambiguous, listeners rely more heavily on their prior expectations to determine the intended meaning.

In turn, a speaker's choice of utterance depends on how likely it is that the listener will recover the intended meaning from it. According to RSA, speakers should avoid ambiguous expressions when describing an unexpected event and instead opt for more explicit, unambiguous forms—such as full noun phrases—to prevent the listener from defaulting to a prior-driven interpretation. The noisy-channel extension further predicts that speakers might even avoid conjoined verb phrases (1a), despite their lack of ambiguity, because in a noisy communication environment listeners might 'repair' the input by supplying a pronoun if doing so leads to a more plausible interpretation.

Both intuition and empirical findings support these predictions. Prior studies (Achimova et al., 2022, 2024; Rygaev et al., under review) show that listeners' expectations shape how they interpret ambiguous pronouns and how prone they are to

misinterpreting zero anaphors under noisy conditions. They also show that speakers use more full noun phrases and fewer zero anaphors when describing surprising events than when describing familiar ones. Achimova et al. (2022) further developed a computational model of referential expression production, offering additional evidence for a Bayesian account of communication.

In this paper, we present a computational RSA model for referential expression interpretation that takes into account event priors and the possibility of noise.

2 Experimental data

Our model builds on data from Rygaev et al. (under review), who investigated utterances of the type shown in (1) across four types of monster interactions: attacking, rock throwing, waving, and jumping over one another. They found that action priors, measured in a separate experiment, quantitatively affect both the production and interpretation of referential expressions. The stronger the patient bias in the prior expectations—that is, the greater the plausibility that the patient would fall—the more likely listeners were to interpret both pronouns and zero anaphors as referring to the patient. Likewise, speakers were more inclined to use full noun phrases and to avoid zero anaphors when referring to the agent.

On the other hand, when referring to the patient, speakers overwhelmingly used noun phrases (over 90% of the time), and no effect of priors was observed. The authors attribute this to a shift in salience: the most salient referent changes from the agent to the patient, prompting a more explicit referential expression regardless of prior expectations.

3 Model

3.1 Vanilla RSA

The standard RSA framework (Franke & Jäger, 2016) models a pragmatic listener L_1 as reasoning about a pragmatic speaker S_1, who in turn reasons about a literal listener L_0. The model assumes a fixed set of communicable world states s, a fixed set of possible utterances u, and a meaning function $[\![u]\!](s)$ that returns 1 if utterance u is true in world state s, otherwise 0. The model also incorporates prior probabilities $P(s)$ for each world state.

$$P_{L_0}(s|u) \propto P(s) \cdot [\![u]\!](s)$$

$$P_{S_1}(u|s) \propto \exp\big(\alpha \cdot U(u,s)\big)$$

$$U(u,s) = logP_{L_0}(s|u) - C(u)$$

$$P_{L_1}(s|u) \propto P(s) \cdot P_{S_1}(u|s)$$

L_0 interprets utterances literally, selecting among compatible world states based on their prior probabilities, whereas L_1 additionally considers how likely S_1 would be to produce a given utterance to convey a particular world state.

S_1 attempts to maximize the utility $U(u,s)$ by balancing the informativeness of the utterance for L_0 against its production cost $C(u)$. The parameter α scales the utility, modulating how likely the utterance with the maximum utility is to be chosen.

3.2 Noisy-channel RSA

The noisy-channel extension of RSA (Bergen & Goodman, 2015) accounts for the possibility that an utterance may be misperceived. Listeners L_0 and L_1 infer the intended utterance u_i based on the perceived utterance u_p, while the speaker S_1 considers that the intended utterance u_i might be misheard as u_p. The noise function $P(u_p|u_i)$ specifies the probability that u_i is (mis)perceived as u_p.

$$P_{L_0}(s|u_p) \propto P(s) \cdot \sum_i P(u_p|u_i) \cdot [[u_i]](s)$$

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

$$P_{L_1}(s|u_p) \propto P(s) \cdot \sum_i P(u_p|u_i) \cdot P_{S_1}(u_i|s)$$

3.3 Our model

Our model is based on the noisy-channel RSA framework but omits a cost function, assuming for simplicity that all utterances carry the same cost. We consider two possible states of affairs: *agent falls* and *patient falls*, and four possible utterances: *zero* (a zero anaphor), *pro* (an ambiguous pronoun), *agent*, and *patient*. The last two correspond to any noun phrases that unambiguously refer to the agent or the patient of the action, respectively.

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

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

$$P_{L_1}(s|u_p, a) \propto P(s,a) \cdot P_{sal}(s) \cdot \sum_i P(u_p|u_i) \cdot P_{S_1}(u_i|s,a)$$

The ambiguous utterance *pro* is compatible with both world states: $[[pro]](agent\ falls) = 1$ and $[[pro]](patient\ falls) = 1$. The other utterances are unambiguous, each compatible with a single world state: $[[agent]](agent\ falls) = 1$, $[[zero]](agent\ falls) = 1$, and $[[patient]](patient\ falls) = 1$, the rest is 0.

Our model incorporates two types of prior probabilities. First, we include action priors $P(s,a)$, representing the likelihood that the patient of an action will fall, which vary depending on the action. Consequently, all formulas now depend on the action parameter a, which serves as the context for interpreting referential expressions. Second, to capture speakers' strong preference for using noun phrases when referring to the patient due to a shift in salience, we introduce a salience factor $P_{sal}(s)$. This represents the prior probability that the speaker will refer to the most salient entity, which in our data is always the agent. We control this factor with a parameter ab (agent bias), setting $P_{sal}(agent\ falls) = ab$ and $P_{sal}(patient\ falls) = 1 - ab$[1].

For the noise function, we assume that under noisy conditions, listeners may confuse zero anaphors with pronouns and vice versa at a fixed error rate e, so that $P(pro|zero) = P(zero|pro) = e$ and $P(zero|zero) = P(pro|pro) = 1 - e$. Assuming noun phrases are always interpreted correctly, we set $P(agent|agent) = P(patient|patient) = 1$[2].

We performed a Bayesian data analysis in WebPPL (Goodman & Stuhlmüller, 2014) using the Markov chain Monte Carlo (MCMC) method (Gilks et al., 1995). By sampling from an uninformative prior parameter value distribution and conditioning on the likelihood of generating the observed data, this method allows us to compute a posterior distribution over the parameter values, estimating which values are most likely to have generated the data.

The mean parameter estimates from the posterior distribution are: $\alpha = 2.2$, $e = 0.14$, $ab = 0.76$. The estimated action priors are: $jump\ over = 0.57$, $wave = 0.66$, $attack = 0.68$, $throw\ rock = 0.73$. Although action priors do not exactly match

[1] Many studies show that listeners form expectations about which referent is likely to be mentioned next (Arnold, 2008; Bott & Solstad, 2023). We therefore encode this parameter on the listener's side as contextually induced salience priors (analogous to action priors). Salient referents can be more easily identified when referred to with a pronoun. Conversely, when speakers intend to refer to a less salient referent, they tend to prefer more explicit forms to avoid a salience-based pronoun interpretation. This accounts for the bias observed on the speaker's side. We acknowledge that, in our implementation, the literal listener may appear overly pragmatic, and alternative approaches could be explored in future work to address this issue.

[2] In a setting without a cost function, a fully rational speaker would invariably prefer a noun phrase, since it is always more informative than a pronoun or a zero anaphor. In the RSA framework, however, the speaker's degree of rationality is governed by the parameter α. As long as α is finite, the model permits the speaker to make subrational choices.

those obtained in the prior experiment by Rygaev et al. (under review), they preserve the same relative ordering, suggesting a monotonic relationship between these priors and the values from the experiment.

We conducted a posterior predictive check by sampling from the posterior predictive distribution, effectively asking the model to simulate the experimental data using the posterior parameter estimates. As shown in Figure 1, the model's posterior predictions closely align with the observed data: the observed values (red dots) lie within the most probable regions of the model's predictions (black curves). The model predicts, in line with the data, that stronger patient bias in the action increases the likelihood that pronouns will be resolved toward the patient (Figure 1, last two columns). The same pattern appears for zero anaphors (Figure 1, first two columns), though with a smaller effect size.

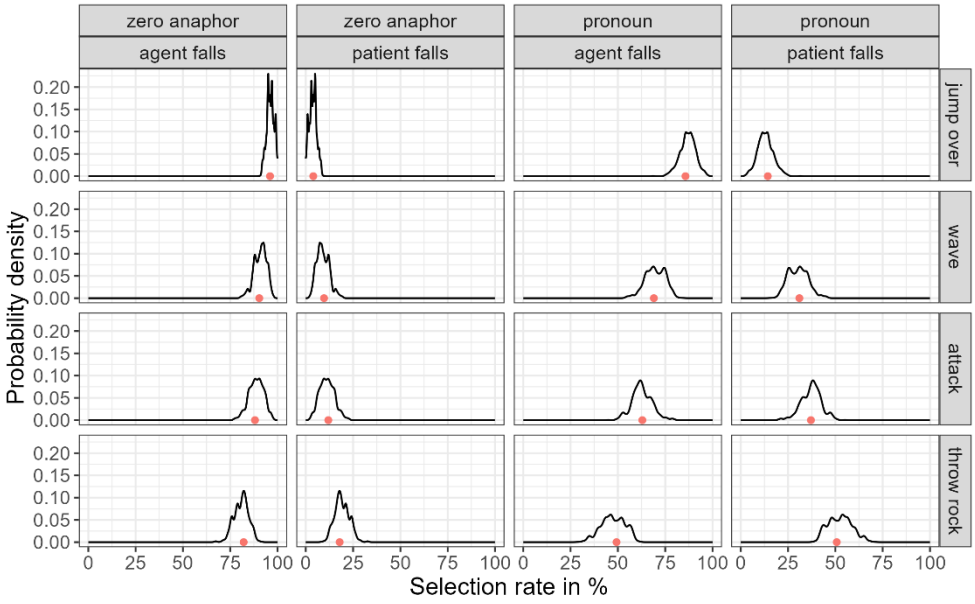


Figure 1: Posterior predictions of the model. Red dots show the observed selection rates for each condition, representing how often participants chose a given outcome for a specific action and expression in the prompt (zero anaphor or pronoun). The black curves depict the model's predicted probability distributions of these selection rates based on the posterior parameter distribution.

3.4 A puzzle

As noted above, speakers overwhelmingly use noun phrases when referring to the patient, consistent with findings from other studies (Demberg et al., 2023). One

might therefore expect that listeners would interpret other forms, such as pronouns, almost exclusively as referring to the agent. Yet, on average, a pronoun was interpreted as referring to the patient in about one-third of cases (Figure 2, right panel). How does our RSA model account for this apparent discrepancy?

The key to the explanation lies in the fact that speakers also rarely use pronouns when referring to the agent (Figure 2, left panel), favoring instead noun phrases or zero anaphors. The overall rate of pronoun usage for agent- and patient-falling outcomes is therefore rather similar, which increases the role of priors in pronoun disambiguation. Since these priors are biased toward the patient-falling interpretation, this accounts for the relatively high rate of pronouns being interpreted as referring to the patient.

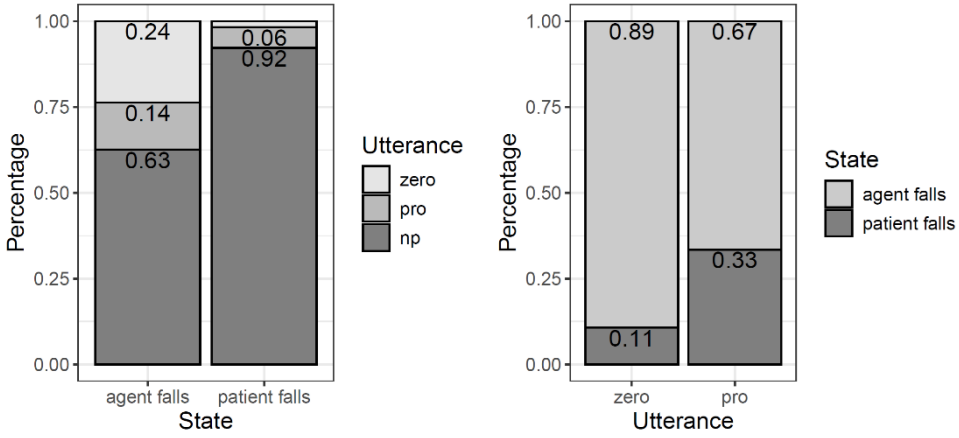


Figure 2: Experimental results. Left: the rate of utterances produced for each state. Right: the rate of states selected for each utterance.

4 Conclusions

We presented an RSA model that accounts for the effects of event predictability and noise in referential expression comprehension. Using Bayesian data analysis with the Markov chain Monte Carlo method, we estimated plausible parameter values and showed that the model's posterior predictions closely match the observed data. We also explained an apparent discrepancy in the experimental data. Our findings support a Bayesian view of language comprehension, where listeners combine prior expectations with linguistic input to infer intended meanings, even when the input is formally unambiguous. They highlight the value of probabilistic models like RSA for capturing pragmatic effects and bridging experimental data with formal theories of language use.

Bibliography

Achimova, Asya, Marjolein van Os, Vera Demberg, and Martin V Butz. “Interpreting implausible event descriptions under noise.” In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46:3399–3406. 2024.

Achimova, Asya, Christian Stegemann-Philipps, Susanne Winkler, and Martin V Butz. “Anaphoric reference in descriptions of surprising events.” In *Proceedings of Sinn und Bedeutung*, 26:20–34. 2022.

Arnold, Jennifer E. "Reference production: Production-internal and addressee-oriented processes." *Language and cognitive processes* 23, no. 4 (2008): 495-527.

Bergen, Leon, and Noah D Goodman. “The strategic use of noise in pragmatic reasoning.” *Topics in cognitive science* 7, no. 2 (2015): 336–350.

Bott, Oliver, and Torgrim Solstad. "The production of referring expressions is influenced by the likelihood of next mention." *Quarterly Journal of Experimental Psychology* 76, no. 10 (2023): 2256-2284.

Demberg, Vera, Ekaterina Kravtchenko, and Jia E Loy. “A systematic evaluation of factors affecting referring expression choice in passage completion tasks.” *Journal of Memory and Language* 130 (2023): 104413.

Franke, Michael, and Gerhard Jäger. “Probabilistic pragmatics, or why Bayes’ rule is probably important for pragmatics.” *Zeitschrift für sprachwissenschaft* 35, no. 1 (2016): 3–44.

Gilks, Walter R, Sylvia Richardson, and David Spiegelhalter. *Markov chain Monte Carlo in practice.* CRC press, 1995.

Goodman, Noah D, and Andreas Stuhlmüller. *The Design and Implementation of Probabilistic Programming Languages.* http://dippl.org. Accessed: 2025-6-24, 2014.

Rygaev, Ivan, Vera Demberg, Martin V Butz, and Asya Achimova. *Event plausibility modulates production and comprehension of referring expressions.* [submitted for publication], 2026.

Winograd, Terry. “Understanding natural language.” Cognitive psychology 3, no. 1 (1972): 1–191.

Authors

Ivan Rygaev, PhD student at the Department of Linguistics of the University of Tübingen