Testing counterintuitive predictions about cost-based inferences in learning from the Rational Speech Act model

Ponrawee Prasertsom (ponrawee.prasertsom@ed.ac.uk)
Centre for Language Evolution, University of Edinburgh

Kenny Smith (kenny.smith@ed.ac.uk)
Centre for Language Evolution, University of Edinburgh

Jennifer Culbertson (jennifer.culbertson@ed.ac.uk)
Centre for Language Evolution, University of Edinburgh

Abstract

The Rational Speech Act (RSA) model has been employed to explain word learning and inferences based on the costs of forms. Here, we focus on a hitherto untested and counterintuitive cost-based effect predicted by RSA: In learning a lexicon with two forms and meanings, learners should prefer an ambiguous costly form over an ambiguous cheap form. We demonstrate this prediction in an RSA model. We then measure reaction times and lexicon ratings in a novel word learning task to test whether a lexicon with an ambiguous costly form is less surprising than one with an ambiguous cheap form. We found no clear evidence for this effect in either measure. We discuss alternative explanations for documented cost-based effects, and the possibility that cost-based inferences may not occur during learning.

Keywords: Rational Speech Act; cost; word learning; pragmatics; language acquisition

Introduction

Rational Speech Act (RSA) models (Frank & Goodman, 2012; Goodman & Frank, 2016), which treat pragmatic effects as arising from recursive reasoning about rational interlocuters' intentions, have been successfully used to explain various linguistic phenomena, most famously implicatures, but also a range of other phenomena (see Degen, 2023 for an overview). RSA has also been employed to study word learning (e.g., Frank, Goodman, & Tenenbaum, 2009; Frank, Goodman, Lai, & Tenenbaum, 2009). Notably, Frank and Goodman (2014) showed that both adult and child participants inferred that a word's meaning corresponds to a distinguishing feature of a referent, as predicted by RSA under the assumption that the rational speaker uses words to be informative. For instance, using the utterance "the dinosaur with a *dax*" for a dinosaur with a bandana, in presence of a dinosaur without, led participants to learn *dax* as a word for 'bandana' as opposed to any other feature shared between the dinosaurs.

A few RSA studies have explored the effect of *cost*, often assumed to correspond to signal length (e.g., Frank & Goodman, 2012; Degen, Franke, & Jager, 2013). RSA models assume that a rational speaker tends to avoid costly forms if other forms are available (although see Baumann et al., 2014 for a nuanced view). This plays a role, for example, in specificity implicatures: Use of a cheap but ambiguous form, *some* 'one or more, possibly all', implies that a cheap unambiguous alternative, *all*, is not true. RSA derives this by assuming that the rational speaker would have used *all* if they meant 'all', and, crucially, preferred the ambiguous but cheap *some* over the unambiguous but costly *some but not all* (Bergen, Levy, & Goodman, 2016; Jäger, 2008). Rohde et al. (2012) used a communication game experiment, with in-game points as cost, to argue that this cost-based inference applies beyond specific words (*some*, *all*), and to referential alternatives more generally. Degen et al. (2013) found that comprehenders' likelihood of making cost-based inferences increases as the cost of the costly form increases, in a task where costs are more naturally operationalised as effort required for typing. Other examples include the derivation of the association between costly forms and improbable meanings in absence of established conventions (Bergen, Goodman, & Levy, 2012), and in both English and novel adverbial intensifiers (Bennett & Goodman, 2018).

Most of the above studies on cost have explicitly modelled communication, but not learning, i.e., where participants try to infer an unknown target lexicon. The artificial intensifier learning task of Bennett and Goodman (2018) did look at the inference of word meaning, but their results may be due to a preference for iconicity in word learning (i.e., mapping unusual meanings to longer forms; Haspelmath, 2008). In other words, while RSA studies exist on both word learning and cost-based inferences, arguably none have zeroed in on the effects of cost *on the learning of the underlying lexicons*, as distinct from the rational agent's behaviour in communication. Here, we focus on a counterintuitive prediction from RSA in this area. In essence, the model predicts that, given a costly and a cheap form, the *costly* form is inferred to be more likely to have *more* meanings than the cheap form. RSA makes this prediction because it is possible for a rational speaker to prefer a cheap form for a given meaning due to low cost, rather than because it is the only way to express the meaning. This leads to the inference that the costly form may also be viable in principle, but not used due to cost. In contrast, if a rational speaker uses a costly form to express a meaning, it must have been used because it is the only way to express that meaning (otherwise the cheap form would have been used). Thus the cheap form is unlikely to share the same meaning. This inference is predicted to be stronger as the cost of the costly form grows. We find this prediction surprising, not least because in natural languages, cheap (i.e., short, easy-to-produce) forms tend to have more meanings (Piantadosi, Tily, & Gibson, 2012).

Below, we implement an RSA model that shows this be-

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

haviour. Then, we present a novel word experiment where we measure reaction-time and metalinguistic lexicon judgments to test whether humans make this kind of cost-based inference, but find no evidence in support of the prediction.

Model

Here we implement a simple RSA model with cost. In the model, a rational agent maintains a distribution over possible lexicons $P(\mathcal{L})$. We define each lexicon $\mathcal{L}$ as a matrix of form f by meaning m, where the value is 1 when f can mean m, and 0 otherwise. The literal listener distribution L_0 is proportional to the prior over meanings $P(m)$ when the form is compatible with the intended meaning.

$$L_0(m \mid f, \mathcal{L}) \propto P(m) \cdot \mathcal{L}[f, m] \tag{1}$$

A level-n speaker S_n assigns probabilities to forms based on how likely the level-$n-1$ listener would be to infer the correct meaning given the form and its cost $c(f)$. λ controls how deterministic S_n will be (higher λ leads to more deterministic behaviour).

$$S_n(f \mid m, \mathcal{L}) \propto \exp(\lambda \cdot (\ln(L_{n-1}(m \mid f, \mathcal{L})) - c(f))) \tag{2}$$

A level-n rational listener L_n assigns probabilities to meanings based on S_n and $P(m)$.

$$L_n(m \mid f, \mathcal{L}) \propto S_n(f \mid m, \mathcal{L}) \cdot P(m) \tag{3}$$

We assume that the goal of lexicon learning is to infer the posterior $P(\mathcal{L} \mid D)$ from data D (form-meaning pairings). Following previous work (e.g., Nedelcu & Smith, 2022; Lewis & Frank, 2013; Frank & Goodman, 2014), an RSA agent learns by updating the posterior based on a rational speaker's distribution over forms. For example, assuming S_1 as a reference for learning, the learning can be formalised as in (4).

$$P(\mathcal{L} \mid D) \propto P(\mathcal{L}) \cdot \prod_{(m,f) \in D} S_1(f \mid m, \mathcal{L}) \tag{4}$$

With this model, we can illustrate the idea outlined in the introduction. We assume 2 forms (mnemonically, *cheap* and *costly*), and 2 meanings ('meaning0' and 'meaning1'). We assume the rational speaker is S_1, $\lambda = 1$, and the priors $P(\mathcal{L})$ and $P(m)$ are uniform. We consider the space of $\mathcal{L}$ to be all 9 possible 2×2 lexicons where every meaning has at least one form that can express it. To test the effect of cost, we manipulate the cost of each form. We varied costs along the log scale to aid interpretation. We fixed the cost of *cheap* at $\ln(1) = 0$ and varied the cost of *costly* from $\ln(0)$ to $\ln(6)$ (1 to 6 times on the linear scale). We consider a learner exposed to data consisting of 5 pairs of (*cheap*, 'meaning0') and 5 pairs of (*costly*, 'meaning1'), i.e. data where 'meaning0' maps consistently to *cheap* and 'meaning1' to *costly*.

Four lexicons are compatible with these data. The first is the ONE-TO-ONE lexicon, where *cheap* exclusively maps to 'meaning0' and *costly* to 'meaning1'. The second is the FREE-FOR-ALL lexicon, where both forms map to both meanings. The GEN-CHEAP lexicon maps *costly* to a single meaning ('meaning1') but treats *cheap* as a general form that maps to both meanings. The GEN-COSTLY lexicon maps the cheap form to a single meaning ('meaning0') but treats *costly* as a general form that maps to both meanings.

Figure 1 shows the posterior $P(\mathcal{L} \mid D)$ after learning under different sets of costs for *cheap* and *costly* for these four lexicons. Other lexicons are assigned a probability of 0 because they could not have generated the data. Unsurprisingly, most probability mass is on the ONE-TO-ONE lexicon across different costs. However, of particular interest is the fact that as the cost of *costly* increases, the GEN-COSTLY lexicon is progressively more likely (from 0.104 to 0.401) and the GEN-CHEAP lexicon less so (from 0.104 to 0.001).

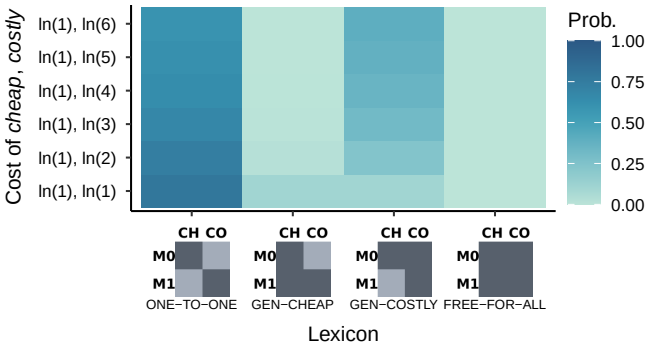

Figure 1: Posterior probability (colour) of data-compatible lexicons (x-axis) given different sets of costs (y-axis). For each matrix (lexicon), *CH* and *CO* are *cheap* and *costly* (the forms); *M0* and *M1* are 'meaning0' and 'meaning1' (the meanings). Dark-grey cells in a matrix are valid form-meaning pairings. Light-grey ones are invalid. The probability of GEN-COSTLY starts out as equal as that of GEN-CHEAP, but increases as the cost of *costly* increases.

This counterintuitive result stems from how S_n is computed. Figure 2 shows the level-1 speaker's distribution S_1 given the ONE-TO-ONE, GEN-CHEAP, GEN-COSTLY and FREE-FOR-ALL lexicon when $c(cheap) = 0$ and $c(costly) = \ln(6)$. To see the effect of cost here, notice that the probability of using *cheap* is not penalised by cost. This means that it may be preferentially used to express 'meaning0' even if *costly* could also be used to express that meaning, as in the GEN-CHEAP lexicon (Figure 2, top right). On the other hand, the S_1 probability of using *costly* is penalised by cost. Thus it can not be the preferred form for an ambiguously expressed meaning. Only if there is no alternative for 'meaning1', will the speaker use *costly*. In other words, *costly* would only be used when *cheap* cannot also mean 'meaning1', as in the ONE-TO-ONE and GEN-COSTLY lexicons (Figure 2, top left and bottom left) but crucially not the GEN-CHEAP lexicon (Figure 2, top right). Thus, the data our learner is exposed to is ambiguous between the ONE-TO-ONE and GEN-COSTLY lexicons, and this ambiguity grows as the cost of *costly* in-

creases. Note, too, that in the FREE-FOR-ALL lexicon, the probability of using either form is inversely proportional to its cost. The fact that the data did not follow this distribution explained why it was assigned almost no probability mass in Figure 1.

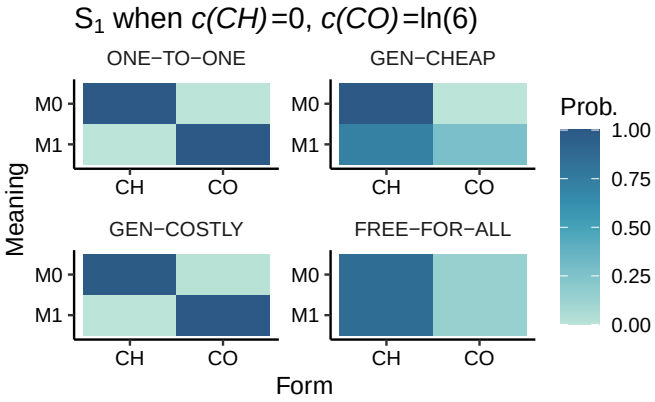

Figure 2: Level-1 speaker distribution S_1 when $c(cheap) = 0$ and $c(costly) = \ln(6)$ given different lexicons (facets). Each row is a distribution over forms (x-axis) given meaning (y-axis). The S_1 behaviour is virtually the same given the GEN-COSTLY or the ONE-TO-ONE lexicon.[1]

In the next section, we test the prediction from this model, namely that if one form is more costly than the other, the lexicon in which the costly form has both meanings will be assigned a higher probability than the one where the cheap form has both meanings.

Lexicon Inference Experiment[2]

Here, we test the broad prediction from the RSA model outlined above: Given data where each form is paired with only one meaning, the model predicts that participants should be more likely to infer a lexicon where the costly form has both meanings (a GEN-COSTLY lexicon) than a lexicon where the cheap form has both meanings (a GEN-CHEAP lexicon). We test this prediction in an artificial lexicon induction task, where participants observe and learn from communicative interaction between two individuals, Jack and Jill. There are two forms and two meanings, and one form is longer, and takes longer for the speaker to produce than the other. Participants observe Jack and Jill's communicative interactions in the context of a messenger-app type interface (as in Figure 3). Jack types messages consisting of one of the two words, and Jill messages back one of the images as her interpretation. On each trial, participants must guess what Jill's response (i.e., the meaning) will be, given a typed word (i.e., the form) from Jack. At the start of the task, the messages back and forth are consistent with a one-to-one lexicon. However, once participants are sufficiently accurate at guessing Jill's response, the behaviour *switches*. If participants, like the RSA model above, have inferred that a GEN-COSTLY lexicon is actually more likely than a GEN-CHEAP lexicon given the data they have observed, then they should be less surprised at a switch that is consistent with this lexicon, and more surprised by a switch to a GEN-CHEAP lexicon. We test whether participants are less surprised by a switch to GEN-COSTLY than GEN-CHEAP using (a) reaction time measurements of participants' guesses, and (b) a metalinguistic judgment task in which participants are asked to rate the GEN-COSTLY and GEN-CHEAP lexicons relative to their ONE-TO-ONE counterparts.

The rationale for using this unusual task was to probe participants' inferences about the underlying lexicon, while not putting participants into a communicative situation themselves, and without inducing inferences that could arise if participants assume that they are explicitly being taught the language. As noted in the introduction, previous work looking at the impact of cost has used a communicative framing, where participants form a new communicative convention through interaction with a partner. For example, in Bergen et al.'s (2012) experiment on the association between improbable meanings and form lengths, participants played a communication game in which participants had to select one of two symbols with different in-game costs to convey a meaning (see Kanwal et al., 2017 for a similar but more naturalistic setup). However, these communicative tasks do not separate inferences about lexicons during learning from pressures operating during use, and mechanisms other than cost-based inferencing could shape the outcome of these games (e.g., using the cheap form first, and switching to something more costly only if communication fails could also yield similar results). Our design avoids this, since participants observe the interaction of others, rather than being directly involved themselves. Our design also avoids putting the participant into an explicit *learner-teacher* relation with the source of their data, as this may trigger reasoning about the speaker/teacher's goal, such as the maximisation of the posterior for the correct lexicon, which differs from RSA-style optimisation. We return to the implications of this design and its relationship to designs used in previous studies in the discussion.

Methods

Participants We collected data from 80 self-reported fluent English speakers on Prolific(40 per each of the two conditions). Participants took about 10±5 minutes to complete the experiment and were compensated £2 for their time.

Materials The model suggests that the effect should be clearest when the cost difference is high (e.g., $\ln(1)$ vs. $\ln(6)$). We picked our two artificial forms accordingly. The costly form *zopekilariki* is 6 times longer (in terms of syllables) than the monosyllabic cheap form *gow*, and took 6 times longer to be "typed" and appear (see Procedure). These forms could be mapped to two meanings, an image of an apple or an orange, according to one of three lexicon types. The ONE-TO-ONE type maps each form to one exclusive meaning. The GEN-CHEAP type maps *gow* to both meanings but

[2]Pre-registered on OSF at https://osf.io/pek4a/?view_only=8425fca9d99a457dabf4a6de203b2b16.

Figure 3: Participants wait for Jack's text to appear (1-2), predict Jill's response (3) and receive feedback (4). Images of apple and orange were from `https://freepik.com` by users `brgfx` and `juicy_fish`, used under their free license.

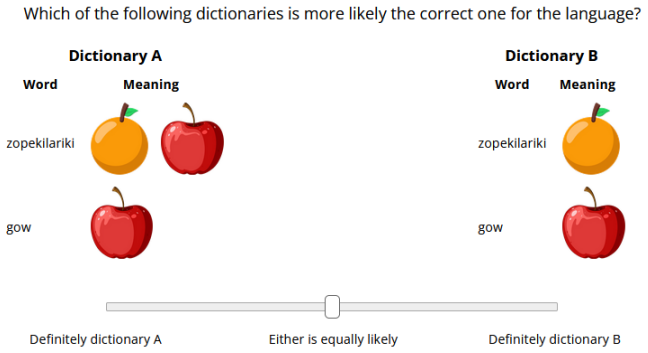

Figure 4: Lexicon rating trial with a GEN-COSTLY lexicon (Dict A) and its ONE-TO-ONE counterpart (Dict B).

zopekilariki to only one. The GEN-COSTLY type maps *zopekilariki* to both meanings, but *gow* to only one. We counterbalanced the mappings across participants; e.g. some saw *gow* associated with apple in the pre-switch trials, others saw *gow* associated with orange.

Procedure At the start of the task, participants were told they would see Jack and Jill playing a game where Jack would type a word and send to Jill, and they had to guess what Jill's image response–the meaning that Jill thought the word meant–would be. Participants were randomly assigned to one of two conditions: GEN-CHEAP or GEN-COSTLY. Participants were then instructed to place their left index finger on the E key and their right index finger on the I key.

In the first ("pre-switch") phase of the experiment, participants saw a messenger-app-like interface. In each trial (Figure 3), a message bubble appeared with a three-dot animation inside, along with the name "Jack" above. After a while, the dots would turn into either *zopekilariki* or *gow*. *zopekilariki* took 6 seconds to appear, whereas *gow* took only 1 second. (i.e., cost is operationalised as length and "typing time"). Participants were then prompted with two images to select (an apple or an orange). A message bubble from Jill would then appear with an image inside. Participants could see and were told through text if their response matched Jill's. We collected the reaction time (RT) for participants' responses in milliseconds (ms), starting when the two image buttons appeared. In this pre-switch phase, a given form typed by Jack always led to the same interpretation by Jill (i.e., the lexicon was ONE-TO-ONE). There were 5 pre-switch trials for each word (10 in total). Participants had to repeat this phase (i.e., all 10 trials) until they reached 80% accuracy (8 out of 10 trials).

After the pre-switch phase, participants proceeded to the post-switch phase without any explicit break or any other indication. Each trial proceeded in exactly the same way, with participants guessing the meaning that Jill would provide to Jack's message. The only difference from the pre-switch phase was the meanings Jill chose for the critical word (*zopekilariki* in the GEN-COSTLY condition, and *gow* in the GEN-CHEAP condition) was mapped to both 'apple' *and* 'orange' (i.e. Jill responded with 'apple' on half of these trials). The mapping of the non-critical word to its single meaning remained the same. In other words, Jill's behaviour switched such that the data was now consistent with the GEN-COSTLY or GEN-COSTLY lexicon, depending on condition. We kept the number of times a word appeared constant across the phases to control for any potential effect of form frequency. Thus, there were 20 trials in this phase: 10 trials for the non-critical word (mapped to its single meaning), and 10 for the critical word (5 mapping to 'apple' and 5 mapping to 'orange'). We ensured that the first trial post-switch presented the critical word with its new meaning. We again collected the reaction time (RT) for participants' responses in milliseconds (ms), starting when the two image buttons appeared.

After these two phases, participants rated the relative likelihood of the pre-switch ONE-TO-ONE and the post-switch GEN lexicon. The lexicons were represented as dictionary tables shown at the two ends of a slider, as in Figure 4. Participants were asked to use the slider to indicate which of the two dictionaries was more likely the correct one for the language.

The experiment was implemented in jsPsych (de Leeuw, Gilbert, & Luchterhandt, 2023). We used R (R Core Team, 2024) with `brms` (Bürkner, 2021) for Bayesian modelling.

Results

In this experiment, participants observed communicative interactions that produced data consistent with a one-to-one lexicon with two meanings expressed by two forms, one cheap and one costly. After they learned these mappings, the lexicon switched such that a critical form—the costly form in the GEN-COSTLY condition, or the cheap form in the GEN-CHEAP condition)—was used for both meanings. We predicted that after this switch, the average RT of trials involving these critical words would be higher in the GEN-CHEAP condition than in the GEN-COSTLY condition, since, according to the RSA model implemented above, the GEN-COSTLY lexicon has higher posterior probability after the pre-switch training data and the switch to the GEN-COSTLY lexicon should therefore be less surprising. Second, we predicted that, in the final lexicon rating trial, participants would rate the GEN-

COSTLY lexicon as more likely than the GEN-CHEAP lexicon, both relative to their pre-switch ONE-TO-ONE lexicon.

Based on visual inspection, we excluded 21 outlier trials with an RT over 10,000 ms or under 100 ms (10 and 11 from the pre- and post-switch phase, respectively). Figure 5 shows the per-trial log-RTs in post-switch trials after exclusion. Contrary to the model's prediction, the average log-RT for the GEN-CHEAP condition was lower by about 0.1 log-ms (6.743 log-ms vs. 6.817 log-ms), or about 1.1 times (50 ms) on the linear scale. We fit a hierarchical Bayesian regression

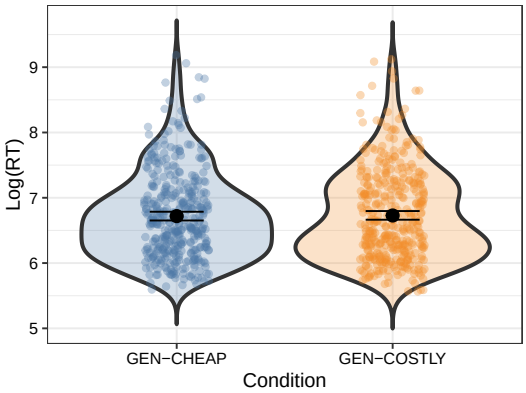

Figure 5: Per-trial reaction times in log-ms (faded dots) in post-switch trials (*y*-axis) by condition (*x*-axis). Error bars are 95% C.I.a round the means (black dots).

model predicting RTs from condition with a by-participant random intercept and a by-item random intercept and random slope on condition, where *item* is defined as the pre-switch form-meaning mapping for a given word (e.g., *gow*-'orange'; included as a control for potential effects of the mapping, which we counterbalanced). We used treatment coding for condition (0 = GEN-CHEAP, 1 = GEN-COSTLY), and a shifted log normal link with the shift parameter α that corresponds to a minimum possible RT. [3] Posterior estimates suggest a highly uncertain, possibly negligible effect with a 95% credible interval including 0 (β_0 = 6.243, 95% CrI [5.172; 7.380], $\beta_{\text{Cond}} = 0.027$, 95% CrI [-1.931; 1.996]; α = 242.002, 95% CrI [229.221; 251.458]). The estimated medians of the effects (expected values of the posterior predictive given different conditions) are almost identical at $\approx$ 912 ms (GEN-CHEAP: 906.343 ms, 95% CrI [471.132; 2344.729] vs. GEN-COSTLY: 918.530 ms, 95% CrI [362.011; 4694.719]).

Note that participants were exposed to more data in the pre-switch phase if they failed to pass the pre-switch 80% accuracy at once. As a result, it is possible that some participants may have assigned a disproportionately high probability to the ONE-TO-ONE lexicon. To explore this, we fit the same model using the participants who passed the threshold in one attempt (32 participants in GEN-CHEAP and 30 in GEN-COSTLY). The results were virtually the same ($\beta_0 = 6.179, 95\%$ CrI [5.030; 7.431]; $\beta_{\text{Cond}} = -0.020$, 95% CrI [-1.941; 2.058]; $\alpha = 239.902$, 95% CrI [225.539; 250.465]).

Next, we considered the lexicon ratings. We normalised the ratings so they fall between 0 and 1, and standardised so that a higher rating means a higher likelihood of a GEN lexicon (GEN-COSTLY or GEN-CHEAP). As Figure 6 shows, the average lexicon rating in GEN-CHEAP (0.594) is slightly higher than in GEN-COSTLY (0.5), again contrary to the RSA model's prediction. To analyse this data, we fit a Bayesian

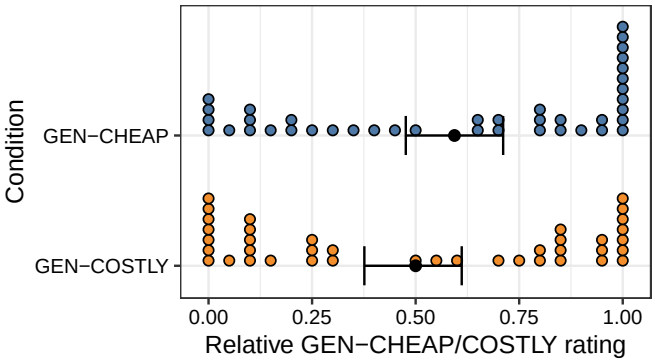

Figure 6: By-participant GEN lexicon ratings (*x*-axis) relative to their ONE-TO-ONE counterparts by condition (*y*-axis). Error bars are 95% C.I. around the means (black dots).

zero-one-inflated Beta regression (Liu & Eugenio, 2018), predicting from condition α (probability of 0 or 1), γ (probability of 1 given 0 or 1), μ (probability over the $(0,1)$ range, exclusive) and ϕ (precision for the Beta distribution).[4] Posterior estimates suggest highly uncertain effects, most of which are possibly non-existent (credible intervals including 0). For the GEN-CHEAP condition, there was a negligible and highly uncertain bias for GEN lexicons (β_0 for μ = 0.036, 95% CrI [-0.390; 0.463]; β_0 for ϕ = 0.703, 95% CrI [0.217; 1.143]), but against extreme ratings (β_0 for $\alpha = -0.501$, 95% CrI [-1.140; 0.115]). However, 1 (GEN) is favoured over 0 (ONE-TO-ONE) when the ratings are extreme (β_0 for γ = 1.048, 95% CrI [-0.064; 2.295]). In the GEN-COSTLY condition, participants were more biased towards ONE-TO-ONE lexicon, but the effects are small and uncertain (β_{Cond} for $\mu = -0.028$, 95% CrI [-0.663; 0.596]; β_{Cond} for ϕ = -0.138, 95% CrI [-0.785; 0.506]; β_{Cond} for α = -0.118, 95% CrI [-1.032; 0.757]; β_{Cond} for γ = -1.089, 95% CrI [-2.781; 0.529]).

Finally, one might wonder whether participants may not have switched to the GEN lexicon due to strong priors for the ONE-TO-ONE lexicon. Such a prior could, for example, a "mutual exclusivity" bias (Clark, 1987; Markman & Wachtel, 1988; Lewis & Frank, 2013). We explored this possibility by looking at how persistently participants continued to

[3] The condition term (GEN-COSTLY's effect relative to the reference) is given a weak $\mathcal{N}(0, 2^2)$ prior (log scale), and a $\mathcal{N}(150, 50^2)$ prior with a lower bound of 0 for α (linear scale), penalising unreasonably low RTs like 50ms. We used default `brms`' priors otherwise.

[4] We expected participants to be biased toward 0 or 1 ratings so we set a *Logistic*$(0.3, 1)$ prior on the intercept for α, slightly favouring 0 and 1, and a *Logistic*$(0, 0.95)$ prior on the intercept for γ, slightly favouring equal chances of 0 and 1. We defaulted to vague priors otherwise.

choose the original meaning for the post-switch trials. We compared the 5 pre-switch trials and the last 5 post-switch trials involving critical words. We found that participants had clearly started selecting the post-switch meaning for the critical word (Pre-switch prop.: 0.923, 95% CI [0.893; 0.948] vs. Post-switch prop.: 0.748, 95% CI [0.708; 0.790]). Of course, this does not rule out that participants still had a prior for ONE-TO-ONE lexicons, but this, together with the rating data, suggests that no very strong prior ruled out the GEN lexicons.

Discussion

This study has focused on a counterintuitive prediction from RSA: That learners should be more likely to infer a lexicon where a costly form is ambiguous rather than a lexicon where a cheap form is ambiguous, all else equal. We first showed that this prediction obtains because a lexicon with an ambiguous costly form (but crucially not one with an ambiguous cheap form) produces the same speaker probability distribution as a lexicon with a one-to-one form-meaning mapping. We then tested this experimentally using a task in which participants had to infer a lexicon based on an observed communicative exchange. At the start those observations were consistent with one-to-one mappings, but switched such that either the costly form or the cheap form because ambiguous. If participants had inferred that a GEN-COSTLY lexicon is nevertheless likely given the initial one-to-one data, we expected the former shift to be less surprising to participants than the latter. We were unable to find experimental evidence in support of the RSA prediction for this kind of cost-based inference. We found no evidence from reaction times (or very weak evidence in the opposite direction) that participants were more surprised if the costly form was generalised than when the cheap form was. Additionally, in a metalinguistic rating task, we found a highly uncertain, small preference for the *cheap* form to have both meanings, rather than the costly form as predicted by the model.

Why was this the case? It is possible that the RSA prediction is correct, but our experimental design was not able to produce the effect. Given that we looked for this effect by measuring reaction time, it is worth noting that RT effects in psycholinguistic studies can be very small, e.g., on the order of 10s of ms (Hamrick, 2022), and the variance of reaction times in our study was high. Thus, the sample size we used may not be sufficient to construct credible intervals that are sufficiently narrow for meaningful inferences. Even if we did have sufficient power, it is also possible that the switch from one-to-one mappings to an ambiguous mapping for one form was unnatural in the absence of a change in cost or some other external stimulus, or simply unclear. An alternative possibility, which we explored above, is that in our participants have a very strong prior for ONE-TO-ONE lexicons. Our analysis in the previous section suggests that any such prior (if it exists) is not absolute, however a strong prior over lexicons may have obscured any difference between our conditions.

As noted above, our somewhat unusual design was motivated by a desire to de-confound lexicon *inference* from other behavioural effects that may exist when participants are themselves reasoning about their own, or their interlocutor's communicative intent. As reviewed above, communication-based demonstrations of cost effects on newly-formed conventions (e.g., as in Bergen et al., 2012; Kanwal et al., 2017) may be explained by an alternative strategy that makes no reference to cost-based inference. Similar effects shown in, e.g., the association between degrees of adverbial intensification and word length (Bennett & Goodman, 2018) could potentially result from iconicity (Dingemanse et al., 2015; see also Lewis & Frank, 2016 for a related phenomenon of matching form-meaning complexity). These results are consistent with an *a priori* or learned association of this kind based on experience with natural lexicons. On the role of cost in *some/all* specificity implicatures, Rohde et al. (2012) and Degen et al. (2013) both showed that in a communication setting, listeners become more likely to map a cheap ambiguous form (e.g., *tree*) to a meaning coassociated with a costly unambiguous form (e.g., *palm tree*) as the cost of this form increases. However, these studies are not targetting lexicon inference, but instead reasoning about use of an established lexicon. Indeed, crucially most previous experimental studies on cost-based inference involve explicit negotiation of conventions during communication, but not in the context of learning a new lexicon. This opens up the possibility that we failed to find such evidence because the RSA model's predictions do not reflect human behaviour during learning.

Conclusion

We presented and tested a counterintuitive prediction from an RSA model that a costly form should be preferably ambiguous over a cheap form in lexicon learning, but found no clear evidence from reaction time and rating measures. We discuss that this could be an experimental design issue, but also argue that the model's prediction may be incorrect, as past demonstrations of cost-based inferences appeared inconclusive, and suggest possible differences in RSA's potential to model learning as opposed to communication.

Acknowledgements

We thank anonymous reviewers for their comments and constructive feedbacks, and Elizabeth Pankratz for advice on Bayesian modelling. Ponrawee Prasertsom's studies are funded by the Anandamahidol Foundation, Thailand.

References

Baumann, P., Clark, B., & Kaufmann, S. (2014). Overspecification and the cost of pragmatic reasoning about referring expressions. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

Bennett, E. D., & Goodman, N. D. (2018). Extremely costly intensifiers are stronger than quite costly ones. *Cognition*, *178*, 147–161.

Bergen, L., Goodman, N., & Levy, R. (2012). That's what she (could have) said: How alternative utterances affect lan-

guage use. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 34).

Bergen, L., Levy, R., & Goodman, N. (2016, May). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, *9*(20), 1–91. doi: 10.3765/sp.9.20

Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, *100*(5), 1–54. doi: 10.18637/jss.v100.i05

Clark, E. V. (1987). The principle of contrast: A constraint on language acquisition. In *Mechanisms of language aquisition.* (pp. 1–33). Lawrence Erlbaum Associates, Inc.

Degen, J. (2023). The rational speech act framework. *Annual Review of Linguistics*, *9*(Volume 9, 2023), 519-540. Retrieved from https://www.annualreviews.org/content/journals/10.1146/annurev-linguistics-031220-010811 doi: https://doi.org/10.1146/annurev-linguistics-031220-010811

Degen, J., Franke, M., & Jager, G. (2013). Cost-based pragmatic inference about referential expressions. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 35).

de Leeuw, J. R., Gilbert, R. A., & Luchterhandt, B. (2023). jspsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software*, *8*(85), 5351. Retrieved from https://doi.org/10.21105/joss.05351 doi: 10.21105/joss.05351

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in cognitive sciences*, *19*(10), 603–615.

Frank, M. C., Goodman, N., Lai, P., & Tenenbaum, J. (2009). Informative communication in word production and word learning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 31).

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Frank, M. C., & Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, *75*, 80-96. Retrieved from https://www.sciencedirect.com/science/article/pii/S0010028514000589 doi: https://doi.org/10.1016/j.cogpsych.2014.08.002

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological science*, *20*(5), 578–585.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, *20*(11), 818–829.

Hamrick, P. (2022). Conducting reaction time research in second language psycholinguistics. In *The routledge handbook of second language acquisition and psycholinguistics* (pp. 150–163). Routledge.

Haspelmath, M. (2008). Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics*, *19*(1), 1–33. doi: 10.1515/COG.2008.001

Jäger, G. (2008). Applications of game theory in linguistics. *Language and Linguistics compass*, *2*(3), 406–421.

Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017). Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, *165*, 45–52.

Lewis, M., & Frank, M. (2013). Modeling disambiguation in word learning via multiple probabilistic constraints. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 35).

Lewis, M., & Frank, M. C. (2016). The length of words reflects their conceptual complexity. *Cognition*, *153*, 182-195. Retrieved from https://www.sciencedirect.com/science/article/pii/S0010027716300919 doi: https://doi.org/10.1016/j.cognition.2016.04.003

Liu, F., & Eugenio, E. C. (2018). A review and comparison of bayesian and likelihood-based inferences in beta regression and zero-or-one-inflated beta regression. *Statistical methods in medical research*, *27*(4), 1024–1044.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*(2), 121-157. Retrieved from https://www.sciencedirect.com/science/article/pii/0010028588900175 doi: https://doi.org/10.1016/0010-0285(88)90017-5

Nedelcu, V. C., & Smith, K. (2022). The complexity of a language is shaped by the communicative needs of its users and by the hierarchical nature of their social inferences. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280-291. Retrieved from https://www.sciencedirect.com/science/article/pii/S0010027711002496 doi: https://doi.org/10.1016/j.cognition.2011.10.004

R Core Team. (2024). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rohde, H., Seyfarth, S., Clark, B., Jäger, G., & Kaufmann, S. (2012). Communicating with cost-based implicature: A game-theoretic approach to ambiguity. In *Proceedings of semdial 2012 (seinedial): The 16th workshop on semantics and pragmatics of dialogue* (pp. 107–116).