Competition from novel features drives scalar inferences in reference games

Jennifer Hu (jennhu@mit.edu)
Brain and Cognitive Sciences
Massachusetts Institute of Technology

Noga Zaslavsky (nogazs@mit.edu)
Brain and Cognitive Sciences
Center for Brains, Minds, and Machines
Massachusetts Institute of Technology

Roger P. Levy (rplevy@mit.edu)
Brain and Cognitive Sciences
Massachusetts Institute of Technology

Abstract

Scalar implicatures, one of the signatures of pragmatic reasoning, are believed to arise from competing alternative utterances, which the listener knows that the speaker could have used to express a strengthened meaning. But do scalar implicatures also arise in the presence of nonce objects, for which no alternative name is known? We conduct a series of experiments assessing the degree of scalar strengthening driven by familiar and nonce objects. We find that nonce objects can derive scalar implicatures as strongly as familiar objects in simple reference games. Our experiments also reveal an asymmetry in the relative strengths of familiar- and nonce-driven inferences: relative to the prior, participants preferentially interpret the name of a shared feature as referring to an object with an additional nonce feature over an object with an additional familiar feature, suggesting that familiar alternatives exert greater scalar pressure than nonce alternatives. We also present exploratory model simulations suggesting that our results may be explained by rationally reasoning about a high-cost description of the novel object. Our findings support the idea that novel lexical entries may be generated from one-shot encounters and spontaneously used in pragmatic inference.

Keywords: scalar implicature, reference games, pragmatics, lexical uncertainty, novel objects, alternatives

Introduction

Humans resolve linguistic ambiguity in remarkably flexible ways. One signature pattern of pragmatic reasoning is **scalar implicature**, which arises when an utterance is interpreted as excluding the meaning of more informative utterances on the same scale (Grice, 1975; Horn, 1972). For example, "some of the students passed the exam" implicates that *not all* students passed the exam, since the speaker could have said "all of the students passed the exam" if that had been the case. This cooperative inference requires listeners to reason about the possible alternatives the speaker could have said, which may arise from entailment relationships (e.g., "some" and "all") or ad-hoc scales constructed from a shared referential context.

Ad-hoc scalar implicatures have been extensively studied in reference game settings (Frank & Goodman, 2012; Vogel, Emilsson, Frank, Jurafsky, & Potts, 2014; Stiller, Goodman, & Frank, 2015; Frank, Emilsson, Peloquin, Goodman, & Potts, 2018). These studies have constructed referential contexts by selecting a base object (e.g., snowman) and overlaying familiar, conceptually related objects (e.g., hat, scarf) in a way that gives rise to an entailment-like relationship between features. Figure 1a-i shows one such context, where two snowmen are wearing hats, and one of these snowmen is additionally wearing a scarf. Given the description "hat", humans have been shown to reliably select the snowman wearing *only* a hat, which is consistent with a pragmatically strengthened interpretation of the utterance. This effect is typically attributed to competition from the alternative "scarf", which could have been used to unambiguously describe the hat-and-scarf snowman if that had been the speaker's intended referent. However, it remains unclear whether similar implicatures arise in contexts where no clear alternative exists, such as when the scarf is replaced by a **nonce object** (Figure 1a-ii).

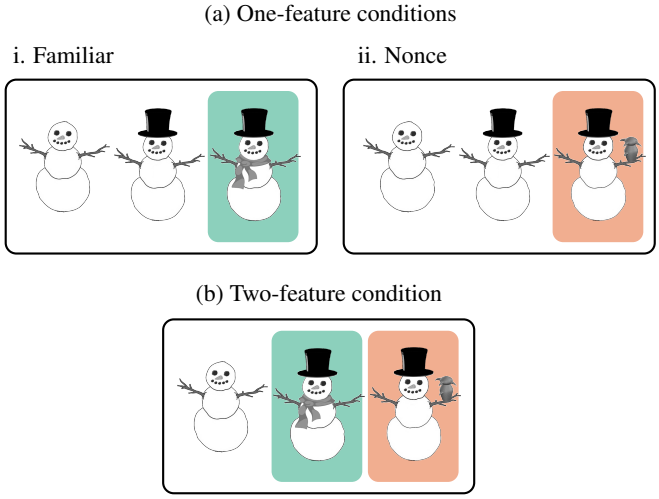

Figure 1: Bob says "hat". In these referential contexts, the snowman you choose is affected by competition from referents with other familiar features (scarf; green shading) or nonce features (greeble; orange shading).

On the one hand, if the label that would be used to indicate a certain feature is not in common ground (Clark & Brennan, 1991) between the speaker and listener, it might not enter into the computations underlying scalar inference. This view would not predict a scalar implicature in contexts like Figure 1a-ii, since there is no competing alternative evoked by the referent with the nonce feature. On the other hand, speakers (children and adults alike) can generate and use a new lexical entry from a single instance of contextually appropriate grounded exposure (Carey, 1978; Markson & Bloom, 1997).

All code and data to reproduce our analyses can be found at https://github.com/jennhu/nonce-SI.

If that ability is used in the machinery of scalar inference, nonce features of potential referents could drive implicature just like familiar features do. Thus, the question of whether scalar implicatures can be driven by nonce objects has implications for how humans spontaneously reason about common ground and the lexicon in previously unobserved contexts.

In this paper, we investigate the following questions: To what extent do nonce objects induce scalar strengthening? And what are the relative strengths of familiar- and nonce-induced scalar implicatures when both features potentially compete as alternatives to a named shared feature? We first conducted an experiment using a basic reference game, and found that nonce objects derive scalar implicatures as strongly as familiar objects. To reinforce the interlocutor's lack of lexical knowledge about the nonce object, we conducted a second experiment with an additional familiarization phase, and found that familiar features exerted greater relative scalar pressure than nonce features. Our results suggest that novel lexical entries may be generated from one-shot encounters and spontaneously used in scalar inference, but these novel alternatives do not drive inferences as strongly when familiar alternatives also compete to a named shared feature.

Related work

Prior work has shown that humans robustly make simple scalar implicatures in reference games where the competing features are familiar and nameable (e.g., Vogel et al., 2014; Frank et al., 2018). It has also been demonstrated that children and adults rationally integrate common ground, informativeness, and speaker history in learning and interpreting novel words (e.g., Markman & Wachtel, 1988; Bohn, Tessler, & Frank, 2019; Bohn, Tessler, Merrick, & Frank, 2020). A crucial difference between these latter studies and ours is that participants in our experiments never encounter a novel word. Instead of identifying the referent of novel words such as "dax", our participants were tasked with identifying the referent of familiar words such as "hat" among a context potentially containing targets with novel features. To the best of our knowledge, it has not been investigated whether scalar inferences are driven by novel objects, and how the degree of such inferences compares to those driven by familiar objects.

Experiment 1

In our first experiment, we tested the strength of scalar implicatures using a basic reference game paradigm, following Frank et al.'s (2018) methodology. All experiments were performed using psiTurk (Gureckis et al., 2016) on Amazon.com's Mechanical Turk (MTurk).

Methods

Procedure Before starting the experiment, participants first saw a page introducing a character Bob, and were told "Bob likes to describe objects with one word." This setup restricted the space of possible alternatives that participants might consider, effectively removing multi-word utterances like "the snowman with only a hat" from consideration. While this allowed us to constrain the alternatives, it also may have decreased the naturalness of the task. We revisit the issue of ecological validity in greater detail in the general discussion.

On each trial, participants saw a set of three referents generated according to one of the three conditions described below (see "Conditions"). The main task prompt read "Bob says [feature]," where the named feature was the one shared between the non-base referents (e.g., "hat" in Figure 1). In order to measure participants' prior preferences for selecting each referent, we also ran a prior elicitation task with the prompt "Bob says ***. Unfortunately, you couldn't hear what he said." For both the main and prior tasks, participants were instructed to click on the object they thought Bob was talking about via 3-alternative forced choice (Figure 2a).

Each participant completed two trials in randomized order: one snowman item, and one tray item (see "Materials"). The condition, task (main vs. prior), and order of referents were randomized on each trial, such that exactly one trial featured a nonce object. After the experiment, participants answered demographic questions, an attention check ("Whom did you meet during this experiment?"), and a question about the nonce object ("What would you call this object?").

Conditions Each set of referents was generated according to one of three conditions. In the one-feature conditions (Figure 1a), participants saw a context with a base object, a base object with one familiar feature (e.g., hat), and a base object with the same familiar feature and one additional feature. This additional feature could either be familiar (e.g., scarf; Figure 1a-i) or novel (e.g., greeble; Figure 1a-ii).

In the two-feature condition (Figure 1b), the context contained a base object, a base object with two familiar features (e.g., hat and scarf; "Familiar+Familiar"), and a base object with one shared familiar feature and one nonce feature (e.g., hat and greeble; "Familiar+Nonce"). In this condition, the two non-base referents are symmetric in logical structure: both have one shared feature and one unique feature.

Materials The items used in our experiment fell into two classes. For continuity with prior work, we first followed the common approach of selecting a base object (snowman) and constructing the other referents by overlaying conceptually related features (hat and scarf). However, this presented a challenge for designing nonce referents. In order to make the nonce feature appear conceptually aligned with the snowman (akin to a hat or scarf), we placed a *greeble* (Gauthier & Tarr, 1997) on the snowman's arm to mimic a perched bird. By placing the nonce feature in a semantically plausible position, however, we risked portraying the nonce object as serving a particular function or being part of a familiar object class.

To address these concerns, we constructed a second class of stimuli using a generic base object with objects in randomized positions (similar to the approach of Asherov, Fox, & Katzir, 2021). For this stimulus class, we used trays as the base objects, with randomly sampled pairs of everyday objects (apple, ball, banana, keys, sunglasses, and teddy bear)

(a) Sample trial

(b) Sample nonce objects

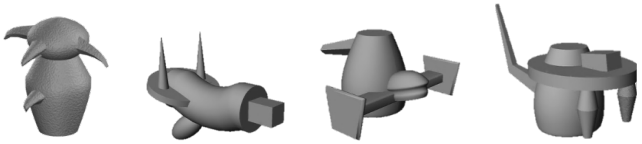

Figure 2: (a) Sample trial in two-feature condition with tray item. (b) Greebles and fribbles were used as nonce objects.

as the familiar features. While we chose to pair greebles with the snowman stimuli for plausibility, the tray stimuli allowed us to use a wider variety of nonce objects with fewer restrictions. We used a set of *fribbles* (Williams, 1998), which are artificially rendered 3D objects with an abstract body and appendage structure. The fribbles are visually similar to the greebles in form and texture (Figure 2b).[1]

All images were presented in grayscale, such that the nonce features could not be easily described with color terms.

Participants We recruited $n = 192$ participants with US-based IP addresses. 24 participants were excluded from our analysis due to not reporting English as their native language and/or not passing the attention check.

Results

Figure 3a shows the results from the one-feature conditions. In both the familiar (green) and nonce (orange) conditions, participants reliably selected the referent consistent with a strong interpretation of the utterance (76% familiar, 85% nonce). Furthermore, after excluding base-referent selections, there was no significant difference between the familiar and nonce conditions (main task $p = 0.55$, prior $p = 0.64$).[2] This suggests that competition from nonce features can derive scalar implicatures as strongly as familiar features.

Figure 3b shows the results from the two-feature condition. The prior is not significantly different from uniform ($p = 0.83$; χ^2 goodness of fit test), and there is no significant difference between selection rates of Familiar+Nonce and Familiar+Familiar across the prior and main task ($p = 1$). This suggests that neither the nonce nor the familiar feature exhibits stronger scalar pressure over the other: upon hearing "hat", participants were equally likely to choose the hat-and-scarf snowman and the hat-and-greeble snowman.

A potential concern is that participants may not be integrating the fact that Bob does not know how to describe the nonce feature. If listeners assume that Bob would use a simple alternative like "alien" to describe the nonce object, then the nonce features would operate like familiar features, which could also explain the symmetry between nonce and familiar features observed in Experiment 1. We sought to rule out this alternate explanation in Experiment 2.

Experiment 2

In a second experiment, we placed Bob's lexical knowledge in the common ground through a familiarization and testing phase. This made participants explicitly aware of the objects that Bob knew and did not know how to name.

Methods

The methods for Experiment 2 were identical to those for Experiment 1, but participants were informed of and tested on Bob's knowledge after viewing the initial instruction screen.

Procedure The conditions and stimuli for the main task were sampled at the beginning of the experiment, producing a set of 4 familiar objects and 1 nonce object that would be seen during the critical trials. During the familiarization phase, participants were shown these objects one at a time. The familiar objects were displayed with the text "Bob says [object name]", and the nonce object was displayed with "Hmm, Bob isn't sure how to describe this object." Participants were only able to advance to the next object after a 1.5 second delay. The 4 familiar objects were presented in randomized order, while the nonce object occurred last to reduce memory demands. During the testing phase, participants were shown all 5 objects in a randomized grid and were asked to click on the ones that Bob knew how to name. Participants then proceeded to the critical trials and postquestionnaire, which included an additional question asking participants to rate how familiar they found the nonce object on a 5-point scale.

Participants We recruited $n = 186$ US-based participants. 56 participants were excluded due to not reporting English as native language, not correctly identifying Bob, making a mistake in the testing phase, and/or rating the nonce object as 3 or higher on the 5-point familiarity scale. This latter exclusion was performed to rule out any participants for whom the greebles or fribbles were not construed as novel objects.

Results

As in Experiment 1, participants reliably selected referents consistent with a strengthened interpretation of the utterance for both familiar and nonce features (Figure 3c), with no significant difference between familiar and nonce one-feature conditions (main task $p = 0.62$, prior $p = 0.13$). Within each one-feature condition, we also investigated whether there

[1] For consistency, we refer to the shared feature as "hat" throughout the paper, but our materials comprise 17 different items.

[2] All reported p-values are given by a two-sided Fisher exact test unless stated otherwise.

were significant differences between Experiments 1 and 2 introduced by the familiarization phase. The difference in priors was not significant between experiments for both one-feature conditions (familiar $p = 0.16$, nonce $p = 0.42$). For the main task, we found no significant difference between experiments for the nonce feature condition ($p = 0.65$), but there was a significant difference for the familiar feature condition ($p = 0.01$).[3] Since Experiment 2 explicitly informed participants that Bob did not know how to describe the nonce object, the similarity between the familiar and nonce conditions cannot be explained by participants assuming there was an easily accessible alternative to identify the nonce referent.

Next, we turn to the two-feature condition (Figure 3d). Relative to the prior (empty bars), participants selected the Familiar+Nonce referent at a significantly higher rate than the Familiar+Familiar referent in the main task (shaded bars) ($p = 0.001$; one-sided Fisher's exact test). This suggests that the Familiar+Familiar referent exerts stronger scalar pressure than the Familiar+Nonce referent: that is, given the ambiguous description "hat", competition from the scarf ultimately wins over competition from the greeble.

Our results across Experiments 1 and 2 provide evidence that nonce features can derive scalar implicatures just as strongly as familiar features in reference game settings. When a referent with an unambiguous familiar feature competes with a referent with an unambiguous nonce feature (as in our two-feature condition), we found that participants selected Familiar+Familiar and Familiar+Nonce at equal rates in Experiment 1. After explicitly informing participants that the interlocutor lacked a description for Familiar+Nonce in Experiment 2, we observed higher rates of selecting Familiar+Nonce (relative to the prior), suggesting that competition from the Familiar+Familiar referent was stronger than competition from Familiar+Nonce.

Role of the prior Since our analysis of the two-feature condition depends on the prior, we wanted to ensure that the observed patterns were not simply artifacts of the prior elicitation method. The "mumble" prior used in both experiments implicitly takes into account Bob's intentions, which potentially interacts with Bob's inability to name certain objects.

We thus repeated Experiment 2 using a prompt designed to *control* the prior. On prior elicitation trials, participants saw "One of the options below has been chosen at random, and Bob needs to describe it to you. Which one do you think it is?". On main task trials, participants saw "One of the options below has been chosen at random, and Bob needs to describe it to you. Bob says [feature]. Click below on the option that you think Bob is talking about." With this method, we could be more confident that any biases reflected participants' baseline sampling of referents and not other factors introduced by the wording of the prompt. We found no significant difference between the "mumble" prior and controlled prior experiments for the main task or prior task. Relative to the prior, participants selected Familiar+Nonce at a marginally higher rate than Familiar+Familiar ($p = 0.05$; one-sided Fisher's exact test). This suggests that our main results in the two-feature condition are not an idiosyncrasy of our prior elicitation method: the relative preference for Familiar+Nonce also holds when we explicitly control the prior.

Finally, we repeated the experiment a third time where participants saw the following prompt on prior elicitation trials: "Bob needs to talk to you about one of the objects below. Which one do you think it is?" The prompt on main task trials was identical to that of the original experiment. We found no significant difference between the "mumble" and "need" experiments for the main task or the one-feature conditions in the prior task, but there was a highly significant difference between the priors for the two-feature condition

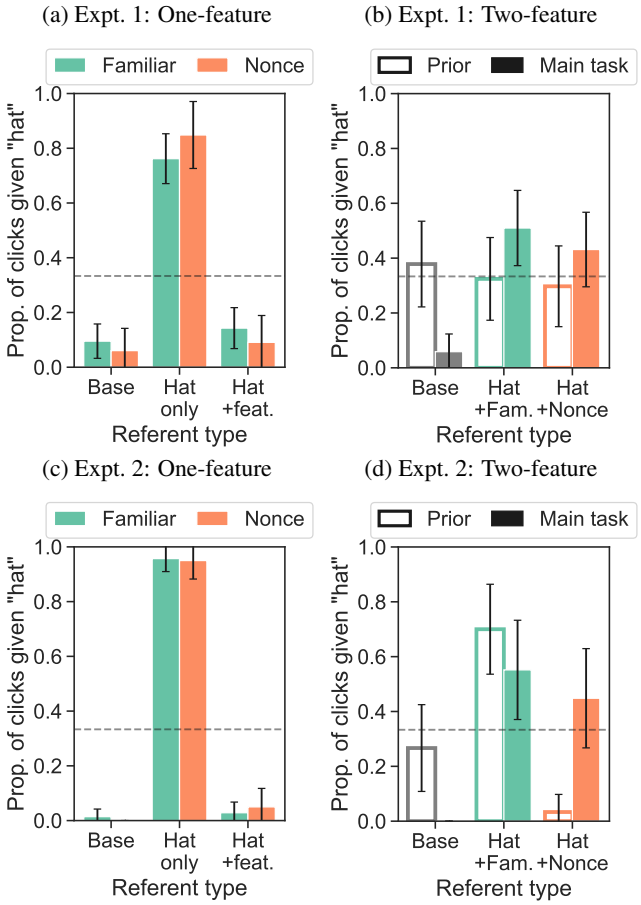

Figure 3: Main results from Experiments 1 (top row) & 2 (bottom row). Error bars indicate 95% binomial CIs. Horizontal dashed lines indicate ⅓ uniform baseline. (a,c) Participants in both experiments reliably chose the referent consistent with a strong interpretation of the utterance, for both familiar and nonce one-feature conditions. (b,d) Participants were ambivalent between Familiar+Familiar and Familiar+Nonce referents in Expt. 1, but chose the Familiar+Nonce referent at a higher rate in Expt. 2 (relative to the prior).

[3] This could be due to increased engagement with the task.

($p = 2.85 \times 10^{-5}$). This reversal of the prior meant that the Familiar+Familiar referent was *more* likely to be selected in the main task than the prior ($p = 0.02$; one-sided Fisher's exact test), in contrast to the "mumble" and controlled prior experiments. Since two of the three priors pattern together – including the controlled prior – we interpret the main task results with respect to the "mumble" and controlled priors, and leave further investigation of different priors to future work.

Potential computational accounts

A natural next question is what computational mechanisms may offer an account for our results. To begin to address this question, we explore two potential mechanisms by building on the Rational Speech Act framework (RSA; Frank & Goodman, 2012), which has been shown to account for many pragmatic phenomena in reference games (Frank et al., 2018).

One potential explanation is that the listener assumes that Bob has a way of uniquely identifying the referent with the nonce feature, but this utterance (which we will call NONCE) may be difficult or otherwise costly to produce. Another possibility is that the listener is uncertain whether Bob knows how to describe the nonce object, and entertains the possibility of multiple lexica that may or may not include an utterance like NONCE. These accounts are neither mutually exclusive nor exhaustive, but provide an intuitive way to begin investigating computational explanations of our empirical findings.

RSA-C: Nonce descriptions are available but costly We start by considering a vanilla RSA model, which formulates pragmatic communication as back-and-forth cooperative reasoning (Grice, 1975). Under RSA, a literal listener L_0 interprets an utterance u by assigning a distribution over meanings based on a lexicon $\mathcal{L}$ and a meaning prior $p(m)$:

$$L_0(m|u) \propto \mathcal{L}(m,u)p(m) \tag{1}$$

Next, given an intended meaning m, a pragmatic speaker chooses an utterance u by soft-maximizing the listener's likelihood while minimizing the cost κ of producing u. The speaker's choice is modulated by a parameter $\alpha > 0$, which is typically interpreted as the speaker's degree of rationality.

$$S_i(u|m) \propto \exp(\alpha(\log L_{i-1}(m|u) - \kappa(u)) \tag{2}$$

Finally, a pragmatic listener defines a distribution over meanings that is Bayesian with respect to the pragmatic speaker.

$$L_i(m|u) \propto S_i(u|m)p(m) \tag{3}$$

The lexicon for the familiar one-feature condition is identical to the one used in many existing studies of scalar implicature in reference games (e.g., Vogel et al., 2014). There are three possible meanings corresponding to the three referents in the context. For each unique feature present in the context, we take there to be a unique utterance in the lexicon (e.g., "snowman", "hat", "scarf").[4] In conditions where there is a referent with a nonce feature, we also take the lexicon to include an utterance NONCE that uniquely identifies that feature. Crucially, κ(NONCE) is greater than the costs of the other utterances, which we take to be equal to each other.

RSA-LU: Nonce objects induce lexical uncertainty The second model is a variant of RSA that incorporates uncertainty about the shared lexicon (LU; Bergen, Levy, & Goodman, 2016). The intuition is that listeners may be uncertain whether Bob knows a way to uniquely identify the nonce object. Under RSA-LU, the pragmatic listener marginalizes over possible lexica $\mathcal{L} \in \Lambda$ (Equation (4)), and the pragmatic speaker and literal listener correspondingly condition on a lexicon $\mathcal{L}$. This model also requires a prior over lexica, $p(\mathcal{L})$.

$$L_i(m|u) \propto p(m) \sum_{\mathcal{L} \in \Lambda} S_i(u|m,\mathcal{L})p(\mathcal{L}) \tag{4}$$

RSA-C and RSA-LU make identical predictions in the familiar one-feature condition, as there are no nonce objects (and thus no LU). For each condition involving nonce features, we consider two lexica, $\mathcal{L}_1$ and $\mathcal{L}_2$. $\mathcal{L}_1$ is the same lexicon used by RSA-C, where there is an utterance NONCE that can uniquely identify the nonce feature. To test the extent to which cost and lexical uncertainty may contribute to explaining our results, we assume that NONCE in the $\mathcal{L}_1$ has the same cost as the other utterances. $\mathcal{L}_2$, on the other hand, does not have an utterance that uniquely identifies the nonce feature, reflecting the view that no label for the nonce feature is in common ground between the speaker and listener.

Model comparison

Next, we explore the degree to which these two models may account for our empirical observations by presenting a set of model simulations that mimic our experimental settings. We tested $\alpha \in \{0.7, 0.9, 1.1, 1.3, 1.5\}$ and listener depths from L_1 to L_5. For simplicity, we set $\kappa(u) = \kappa_0 = 1$ for all $u \neq$ NONCE, and took a uniform prior over meanings and lexica. While our empirically measured priors are generally not uniform, we emphasize that using a non-uniform prior does not qualitatively change our conclusions about the models – the relationships described below hold for any prior.

As expected, both models predict scalar strengthening in the familiar one-feature condition (Figures 4a and 4b, green bars), consistent with human behavior and prior work. Turning to the nonce one-feature condition, both models can predict the interpretation of "hat" to fall anywhere between fully strengthened (i.e., matching the green bar) to non-pragmatic (i.e., matching the prior). These predictions can be manipulated parametrically by tweaking κ(NONCE) in the RSA-C model, and the lexicon prior in the RSA-LU model. For example, Figure 4a illustrates that RSA-LU predicts the strong scalar implicature pattern revealed by our experiments for a uniform prior over lexica ($\alpha = 1.3$, listener depth 3).[5] Increasing $p(\mathcal{L}_1)$ would push the nonce posterior (orange bars)

[4] While this is an oversimplification of the full range of possible utterances, we believe this modeling choice is justifiable given that participants were told Bob only communicated using single words.

[5] We illustrate results for these parameters because they fall in the range typically reported in empirical studies of reference games, but the general patterns hold across a range of α values.

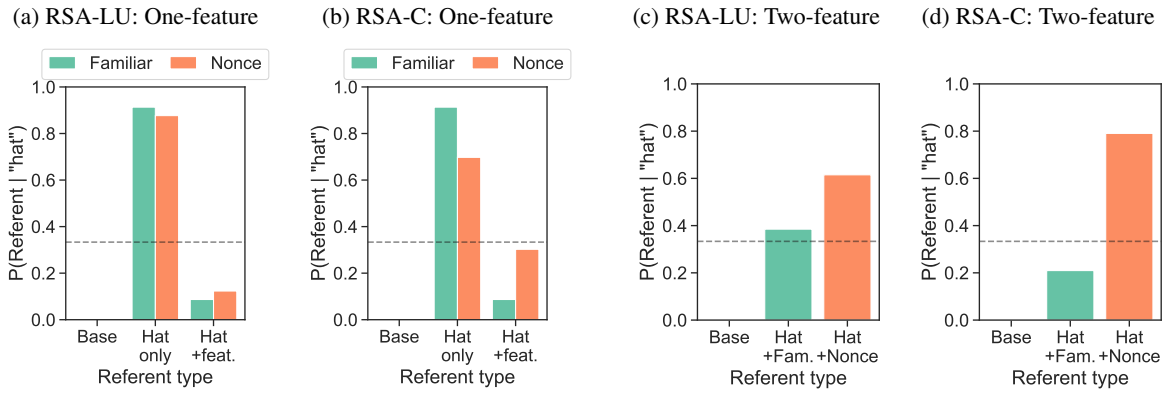

Figure 4: Model listener predictions with uniform meaning prior, uniform lexicon prior, $\alpha = 1.3$, listener depth 3, $\kappa_0 = 1$, and $\kappa(\text{NONCE}) = 2$. While our empirically measured meaning priors are generally not uniform, the relationships between the prior, κ, and listener posterior distributions hold for any prior (see main text).

towards the familiar posterior (green bars), whereas increasing $p(\mathcal{L}_2)$ would push the nonce posterior towards the prior. For the same parameters and $\kappa(\text{NONCE}) = 2$, Figure 4b shows that RSA-C also predicts pragmatic strengthening of "hat", though to a lesser extent than in the familiar condition. Equations (2) and (3) demonstrate that this effect can be modulated through κ under RSA-C: as $\kappa(\text{NONCE})$ approaches infinity, $L(\cdot|\text{"hat"})$ approaches the prior, and as $\kappa(\text{NONCE})$ approaches κ_0, $L(\cdot|\text{"hat"})$ approaches the peaked distribution in the familiar condition.

A similar pattern holds in the two-feature condition. For a uniform lexicon prior, RSA-LU predicts $L(\text{Fam.+Nonce}|\text{"hat"}) > L(\text{Fam.+Fam.}|\text{"hat"})$ (Figure 4c), and the magnitude of this difference increases with $p(\mathcal{L}_2)$. Under RSA-C, the high cost of NONCE weakens its viability as an alternative to "hat", increasing $L(\text{Fam.+Nonce}|\text{"hat"})$ (Figure 4d). This probability decreases with $\kappa(\text{NONCE})$, reaching the prior when all costs are equal.

This initial analysis suggests that both RSA-C and RSA-LU can in principle explain our empirical findings. In order to adjudicate between the two models, recall that in Experiment 2, participants were explicitly told that Bob was unsure of how to describe the nonce object. We expect that this had the effect of reducing lexical uncertainty by increasing the prior probability of $\mathcal{L}_2$ under the lexicon distribution. However, if the RSA-LU model favors $\mathcal{L}_2$, then it would be challenged by our results in the nonce one-feature condition, where we consistently observed that nonce objects drove scalar inferences as robustly as familiar objects. We take this to suggest that RSA-C may be a preferred explanation of our results, which also accords with evidence that speakers can generate and use novel lexical entries from a single exposure (Carey, 1978).

Discussion

We conducted two experiments assessing the degree of scalar strengthening induced by familiar and nonce objects. The basic reference game paradigm in Experiment 1 revealed no significant differences between implicatures derived by familiar and nonce objects. When participants were explicitly informed that the interlocutor did not know how to describe the nonce object in Experiment 2, we observed again that familiar and nonce objects derived scalar implicatures at similar rates. This is consistent with the hypothesis that novel lexical entries may be generated from one-shot encounters and spontaneously used in pragmatic inference. The two-feature condition revealed an asymmetry in the relative strengths of familiar- and nonce-driven inferences: upon hearing the name of a shared feature like "hat", participants selected the hat-and-nonce referent at a higher rate than the hat-and-scarf referent (relative to the prior), suggesting that familiar alternatives exert greater scalar pressure than nonce alternatives.

One issue with this general experimental paradigm is ecological validity. Since pragmatics is precisely concerned with the influence of contextual factors on language use, care must be taken when drawing generalizations from tightly controlled experimental settings to linguistic interactions in the wild. Nevertheless, we believe that this broader class of reference game experiments (conducted not only here but also in many other studies) reveals the possibility of certain types of inferences, which may underlie naturalistic communication under the right conditions. We acknowledge the limitations of this approach and advocate for the development of methods that embed signaling games in more naturalistic settings while maintaining fine-grained experimental control.

In future research, we would like to gain a better understanding of the effect of prior elicitation methods and stimulus types. We also aim to expand our exploratory simulations to a wider range of models. For example, while our analysis has taken alternatives to be realized as strings, other proposals suggest that alternatives operate at the level of *concepts* instead of linguistic forms (Buccola, Križ, & Chemla, 2018), in which case we might expect the cost structure to be symmetric across the nonce and familiar features. Further work is needed to develop a computational account of how alternatives, utterance costs, lexical uncertainty give rise to such flexible patterns of pragmatic reasoning.

Acknowledgments

We would like to thank members of the MIT Computational Psycholinguistics Lab, Harvard Language and Cognition, and three anonymous reviewers for their helpful comments. J.H. was supported by an NSF Graduate Research Fellowship. N.Z. was supported by a BCS Fellowship in Computation. RPL acknowledges support from NSF grants BCS-1551866 and BCS-1456081, a Google Faculty Research Award, Elemental Cognition, and the MIT Quest for Intelligence.

References

Asherov, D., Fox, D., & Katzir, R. (2021). On the irrelevance of contextually given states for the computation of scalar implicatures. In *Proceedings of the Linguistic Society of America.*

Bergen, L., Levy, R., & Goodman, N. D. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, *9*.

Bohn, M., Tessler, M. H., & Frank, M. C. (2019). Integrating common ground and informativeness in pragmatic word learning. In *Proceedings of the Cognitive Science Society.*

Bohn, M., Tessler, M. H., Merrick, M., & Frank, M. C. (2020). *How young children integrate information sources to infer the meaning of words.* Retrieved from https://psyarxiv.com/2wgfb

Buccola, B., Križ, M., & Chemla, E. (2018). *Conceptual alternatives: Competition in language and beyond.* Retrieved from http://ling.auf.net/lingbuzz/003208/current.pdf

Carey, S. (1978). The child as word learner. In *Linguistic theory and psychological reality.*

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In *Perspectives on socially shared cognition.*

Frank, M. C., Emilsson, A. G., Peloquin, B., Goodman, N. D., & Potts, C. (2018). *Rational speech act models of pragmatic reasoning in reference games.* Retrieved from https://psyarxiv.com/f9y6b/

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

Gauthier, I., & Tarr, M. J. (1997). Becoming a "Greeble" expert: Exploring mechanisms for face recognition. *Vision Research*, *37*(12), 1673 – 1682.

Grice, H. P. (1975). Logic and conversation. In *Syntax and Semantics: Speech Acts.*

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., ... Chan, P. (2016). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, *48*(3), 829–842.

Horn, L. R. (1972). *On the semantic properties of logical operators in English*. PhD Thesis, UCLA.

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.

Markson, L., & Bloom, P. (1997). Evidence against a dedicated system for word learning in children. *Nature*.

Stiller, A. J., Goodman, N. D., & Frank, M. C. (2015). Ad-hoc implicature in preschool children. *Language Learning and Development*, *11*(2), 176–190. Retrieved from https://doi.org/10.1080/15475441.2014.927328

Vogel, A., Emilsson, A. G., Frank, M. C., Jurafsky, D., & Potts, C. (2014). Learning to reason pragmatically with cognitive limitations. In *Proceedings of the Cognitive Science Society.*

Williams, P. (1998). *Representational organization of multiple exemplars of object categories.* (Stimulus images courtesy of http://www.tarrlab.org/)