That's what she (could have) said: How alternative utterances affect language use

Leon Bergen (bergen@mit.edu)[1], **Noah D. Goodman (ngoodman@stanford.edu)**[2], **Roger Levy (rlevy@ucsd.edu)**[3]
[1]Department of Brain and Cognitive Sciences, MIT, Cambridge MA 02139,
[2]Department of Psychology, Stanford University, Stanford CA 94305,
[3]Department of Linguistics, UC San Diego, La Jolla CA 92093

Abstract

We investigate the effects of alternative utterances on pragmatic interpretation of language. We focus on two specific cases: specificity implicatures (less specific utterances imply the negation of more specific utterances) and Horn implicatures (more complex utterances are assigned to less likely meanings). We present models of these phenomena in terms of recursive social reasoning. Our most sophisticated model is not only able to handle specificity implicature but is also the first formal account of Horn implicatures that correctly predicts human behavior in signaling games with no prior conventions, without appeal to specialized equilibrium selection criteria. Two experiments provide evidence that these implicatures are generated in the absence of prior linguistic conventions or language evolution. Taken together, our modeling and experimental results suggest that the pragmatic effects of alternative utterances can be driven by cooperative social reasoning.

Keywords: Pragmatics; Communication; Bayesian modeling

Introduction

A central observation in the field of pragmatics is that alternative utterances affect our interpretation of language. If the teacher says, "Some of the students passed the test", then this means that not all of them passed, because the teacher would have said so if they did. If someone is asked what they ate at the restaurant, and they say "salad", then this means that they did not also get the lobster; otherwise they would have said so. If someone says, "I got the car to turn on," then this means that turning on the car involved something more unusual than just turning the key. If it hadn't, they could have just said, "I turned on the car." Many other cases like this were described in Grice's (1975) classic.

Horn (1984) proposed a unified account of these disparate cases in terms of his Q and R-Principles. These principles describe how conversational partners are expected to communicate with each other. The Q-Principle states: Say as much as you can. The R-Principle states: Say no more than you must. These principles explain how counterfactual utterances like the ones above have their effect on meaning. If the speaker behaves according to the Q-Principle, then when she says that some of the students passed the test, this must mean that she said all that she could. In particular, she must not have been in a position to say that all of the students passed. Similarly, if the speaker is following the R-Principle, then when she reports that she got the car to turn on, this means that a simpler utterance would not have sufficed to convey her meaning. In particular, simply saying that she turned on the car would not have conveyed her meaning.

A basic question about these principles (or Grice's related maxims of conversation) is the extent to which they capture people's online reasoning when they pragmatically interpret language. The alternative is that such explanations merely provide a succinct way of summarizing pragmatic phenomena, and that pragmatic meanings are learned as part of the grammar, i.e. conventionalized as part of the language. Researchers have argued that some types of pragmatic meanings are computed from the grammar, and not from cooperative social reasoning (Chierchia, 2004). Intuitively, however, some pragmatic inferences generalize to settings in which they could not have been previously learned. Consider the salad/lobster inference described above. This inference is highly context-dependent, and must require a reasoning process that extends beyond what has been learned in the grammar. But then where is the boundary between conventionalized and socially-derived implicatures?

Here we investigate these questions using experiments and computational modeling. Despite the apparent simplicity of explanations in terms of the Q and R-principles, it has been notoriously difficult to develop a formal framework that captures these principles (or the maxims of conversation). In addition, there has been little empirical work investigating whether pragmatic inferences rely on conventionalized meanings. We will be looking at the minimal ways for contrasts between alternative utterances to drive pragmatic interpretation. Specifically, we will be looking at cases in which there are no linguistic conventions whatsoever. If people's pragmatic interpretations show the same sensitivity to contrast in these settings, it will provide evidence that social reasoning explains more, rather than less, of their pragmatic abilities.

We focus on two traditional examples of counterfactual reasoning in pragmatics. The first, *scalar implicatures*, arise because of the contrast between words that fall on an increasing scale of informativeness. The less informative meaning is typically strengthened to the complement of the more informative meaning, as in the case of "some" vs. "all" above. Because the term "scalar implicature" is sometimes reserved to refer to cases where lexical items fall on a canonical scale of informativeness, we will use the term *specificity implicatures* to describe the strengthening of less informative meanings even in the absence of such a canonical scale. The second, which we will call *Horn implicatures*, guides the interpretation of utterances that differ in their complexity (Horn, 1984). Typically, more complex constructions receive marked (or less probable) interpretations. The car case above is an example of a Horn implicature, assuming (as is plausible) that the two expressions have the same literal content.

These two kinds of implicatures will allow us to explore pragmatic contrast effects along two distinct dimensions: informativeness and cost. While we will model both of these ef-

fects as recursive social reasoning, it will emerge that the simplest model of such reasoning that can account for specificity implicatures is insufficient to explain Horn implicatures. We begin with the simpler version of the model and later enhance the model to account for both kinds of effects.

Specificity Implicatures

The Gricean tradition views pragmatics as a special domain of social cognition. Pragmatics, on this approach, is the study of social agents who want to cooperate with each other to exchange information. Pragmatic phenomena arise as a result of these goals and the agents' reasoning about each other.

Here we develop a model of ideal discourse between two rational agents, a speaker and listener, each with distinct social goals. The speaker wants to communicate a specific meaning to the listener, requiring her to reason about how the listener will interpret her possible utterances. The listener, in turn, wants to determine what meaning the speaker intended to convey, requiring him to reason about which meaning would have led the speaker to send her utterance. The speaker and listener are modeling each other; crucially, the listener takes into account the speaker modeling him, and the speaker takes *this* into account, and so on. In other words, the speaker and listener have *common knowledge* of each others' communicative goals (Lewis, 1969; Clark, 1996).

This recursive social reasoning bottoms out when the listener stops reasoning about the speaker's intentions. In this *base case*, the listener uses his knowledge of the language's semantics or contextual iconicity to interpret the utterance.

We now turn to the formal specification of the model. The literal content of the utterances is specified by a lexicon $\mathcal{L}$, which maps each utterance to a truth function on meanings. If an utterance has no conventional or iconic meaning, then it is given the all-*true* function (i.e. it is a tautology). The listener has a prior distribution P over meanings; in the base case, the listener uses Bayesian inference to update her belief about the intended meaning given the utterance's literal meaning. More precisely, the listener conditions the prior distribution on the utterance being true, essentially filtering P through the literal meaning, leading to a new distribution L_0 with support only on meanings that are consistent with the utterance. That is:

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

where m is a meaning, u is the utterance sent by the speaker, $\mathcal{L}_u$ is a function from meanings to $\{0,1\}$, with $\mathcal{L}_u(m) = 1$ if m is in the denotation of u, and P is the listener's prior distribution over meanings.

Social reasoning enters the model through a pair of recursive formulas that describe how the speaker and listener reason about each other. The formulas describe Bayesian agents S_n and L_n of increasing sophistication. The least sophisticated speaker S_1 reasons about the base, "literal" listener L_0; a slightly more sophisticated listener L_2 reasons about this speaker; and so on. The speaker S_n has a utility function U_n that simultaneously accounts for how informative an utterance is for listener L_{n-1} as well as for its complexity or cost. This is intended to capture both the Q and R-Principles. The speaker's choice of utterance is determined by a softmax decision rule that describes an approximately optimal Bayesian decision-maker (Sutton & Barto, 1998). The listener L_n interprets an utterance by using Bayesian inference to integrate his prior expectations over meanings, given by P, with his model of S_{n-1}, which determines how likely the speaker would have been to use that utterance given each possible meaning.

This recursive model is defined as follows. The speaker's conditional distribution over utterances given interpretations is defined as

$$S_n(u|m) \propto e^{\lambda U_n(u|m)}, \tag{2}$$

where $\lambda > 0$ is the gain on the speaker's softmax decision rule. $U_n(u|m)$ is the speaker's expected utility from uttering u to convey m, defined as

$$U_n(u|m) = \log(L_{n-1}(m|u)) - c(u). \tag{3}$$

Here $c(u)$ is the cost of uttering u (in, e.g., time and effort); the other term measures the communicative benefit of u as the number of bits of information remaining between the listener's posterior distribution $L_{n-1}(m|u)$ and the true meaning m (Frank, Goodman, Lai, & Tenenbaum, 2009). Substituting equation 3 into equation 2, we see that

$$S_n(u|m) \propto (L_{n-1}(m|u)e^{c(u)})^{\lambda}. \tag{4}$$

Hence the speaker prefers low-cost utterances and also prefers to choose an utterance more as the listener is more likely to pick the correct meaning given the utterance. The listener's higher-order interpretations are simply defined as

$$L_n(m|u) \propto P(m)S_{n-1}(u|m). \tag{5}$$

The model defined here is very similar to the iterated best response model of (Jäger & Ebert, 2009).

In situations where the speaker has a choice between utterances one of whose literal meanings is a subset of the other, this model induces an inference that the utterance with the broader meaning should be interpreted as indicating that the narrower meaning does not hold—which we term a specificity implicature. To see why, suppose that there are two possible meanings, *pyramid* and *cube*, and two utterances, "pyramid" and "shape", both of which have equal cost. The literal listener interprets "pyramid" as meaning *pyramid* with probability 1, due to the truth-conditional component of literal interpretation, and interprets "shape" as meaning either *pyramid* or *cube* with probabilities based on the prior, P(pyramid) and P(cube) respectively. For the speaker S_1 reasoning about the literal listener, conveying *pyramid* with "pyramid" has higher utility than conveying it with "shape", since the former term ensures the proper interpretation. S_1 is thus more likely to say "pyramid" than "shape" when she means to convey *pyramid*; and she will obligatorily say "shape" when she means to convey *cube*. The more sophisticated listener L_2 uses S_1's distributions rather than literal meaning, and thus prefers to interpret "shape" as *cube* rather than *pyramid*, since

the likelihood of "shape" is greater when *cube* is to be conveyed than when *pyramid* is. As the speaker and listener reach higher levels of recursive reasoning, these tendencies to say "pyramid" for *pyramid* and to interpret "shape" as *cube* continue to strengthen, both ultimately asymptoting at probability 1 (Figure 2, pink bars; the asymptotes do not depend on λ or P, so long as $\lambda > 1$ and $P(m) > 0$ for all m).

Experiment 1

Experiment 1 investigated whether people will draw specificity implicatures in a novel communicative setting. We investigated this by looking at the simplest setting in which specificity implicatures are possible, a language with only two messages and two meanings. By varying the (non-conventional) semantic content of the messages, we can determine whether competing messages influence interpretations here as they do in richer, conventionalized settings.

We presented people with a simple communication game, which they played with a partner. In the game, one player was the "speaker", who had a specific meaning to communicate, and one player was the "listener", who had to infer this meaning based on the message sent by the speaker. The meaning for the speaker to communicate was randomly chosen to be either a pyramid or a cube. The speaker had the choice between two messages to send the listener: a shape with an iconic relationship with one of the meanings (a triangle for the pyramid), or an "alien" symbol with no obvious connection to either meaning (see Figure 1). If people's reasoning about semantic competition extends to novel settings, then the alien symbol will get a strengthened interpretation: it should be interpreted as the cube, which is the meaning that the speaker could not directly pick out. Likewise, the speaker will recognize that choosing the alien symbol is more likely to communicate the meaning for which there is no iconic symbol available. We thus predict that the speaker will use the alien symbol to communicate this meaning.

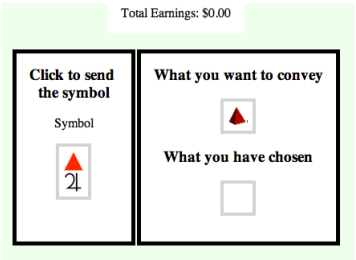


Figure 1: Experiment 1 game interface for speaker.

Methods:

We recruited 40 participants from Amazon Mechanical Turk. They were paid a small amount of money for participation, in addition to performance-based bonuses, as described below.

Participants were told that they had arrived on an alien planet that contained two objects, a pyramid and a cube. Their goal was to successfully play a communication game with a partner, given these objects and two messages that they could send, a triangle and an alien symbol. Participants received 10 rounds of practice as the speaker in order to familiarize with the interface; however, they did not play these rounds with a partner or receive any feedback.

The communication game consisted of five rounds. In each round, participants were randomly assigned a partner and a role as either the speaker or listener. Between rounds, participants were told that they were being randomly matched with a partner. Participants were never identified to each other. The speaker was shown a randomly chosen object that needed to be communicated, and given the choice of the two messages to send. Once the speaker clicked on a message, it was sent to the listener, who was asked to determine what meaning was intended. The listener was given the choice of the two objects; once the listener clicked on one of these objects, the speaker and listener were informed whether their communication was successful. If they were successful, they each received a small bonus payment of $0.06 for that round.

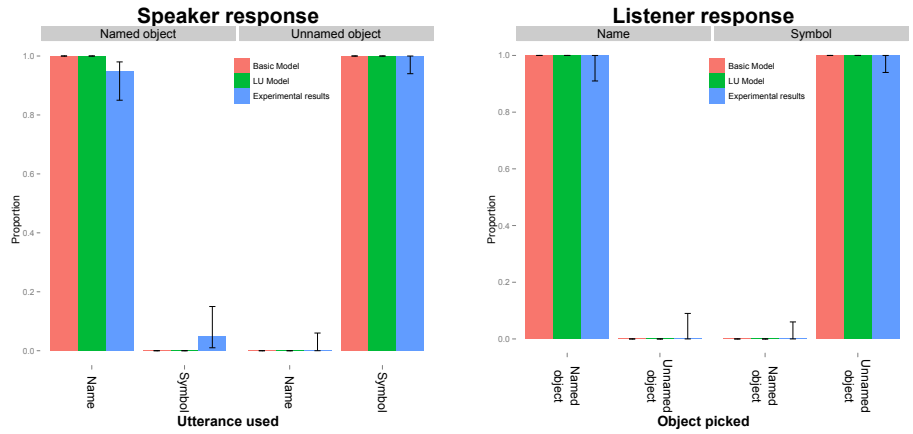
Results and Discussion

There were two questions of interest in this experiment. The first was whether listeners would interpret the alien symbol as the unnamed object (i.e. the object without an iconic message), i.e. the cube. On every trial, the listener interpreted the alien symbol as the unnamed object and the iconic symbol as the corresponding object. The second question was whether the speaker would choose the alien symbol to convey the unnamed object. Participants selected the alien symbol on every trial on which they needed to communicate the unnamed object; and they selected the name on all but two trials on which they needed to communicate the named object. These results are shown in Figure 2. (The displayed model predictions were not sensitive to the value of the model parameters.)

These results provide evidence that participants were sensitive to the semantic contrast between available utterances. Listeners inferred that the speaker would have only used the alien symbol if she needed to communicate the unnamed object. Speakers similarly inferred that the listener would interpret the alien symbol as the unnamed object, and only chose the alien symbol in order to communicate this object.

Horn's Principle

HORN'S PRINCIPLE describes the effects of lexical competition when utterances differ in cost instead of semantic content. The principle states that phrases that are "costlier"—e.g., longer, or involving less-frequent subexpressions—are associated with less probable meanings. For example, *I turned on the car* and *I got the car to turn on* have approximately identical literal meaning; but most speakers would use the shorter sentence to refer to the typical turning of a car key and the longer sentence to some less typical manner of turning on the car. This is the efficient mapping between form and meaning; Horn (1984) and others have documented many instances of such efficient mappings in language.



(a) (b)

Figure 2: Experiment 1 results and model predictions. (a) Speaker responses given the goal of communicating the named object (left) or the unnamed object (right). The y-axis is the proportion of trials on which the speaker chose each utterance. Error bars are 95% confidence intervals. (b) Listener responses when the speaker chose either the name (left) or the alien symbol (right). The y-axis is the proportion of trials on which the listener chose each object.

Perhaps surprisingly, however, the model of intuitive cooperative communication we have introduced fails to predict Horn's principle. Consider the problem of a one-shot speaker-listener signaling game with two utterances, "expensive" and "cheap" (the costs of these utterances reflect their names), and two meanings, *likely* and *unlikely*; nothing distinguishes the utterances other than their cost, and each has the all-*true* meaning. The literal listener L_0 interprets both utterances identically, matching the prior probabilities of the meanings. L_0's interpretation thus provide no information with which the speaker S_1 can distinguish among the utterances; the only thing distinguishing the utterances' utility is their cost. This leads to an across-the-board dispreference on the part of S_1 for "expensive", but gives no starting point for more sophisticated listeners or speakers to break the symmetry.

In fact, Horn's principle has been extraordinarily difficult to derive within other formal frameworks as well. The problem we posed is equivalent to the multiple equilibrium problem for signaling games in economics, which has been investigated for the last 30 years (Cho & Kreps, 1987; Chen, Kartik, & Sobel, 2008). The fundamental difficulty involves ruling out inefficient equilibria, e.g. those in which "expensive" is associated with *likely* and "cheap" with *unlikely*, in the absence of prior conventions or ad-hoc rules for choosing among equilibria. Some recent work by linguists (van Rooij, 2004, 2008; Franke, 2009) has attempted to derive Horn's principle by using evolutionary game theory or hybrid game-theoretic models, but none has found a derivation for one-shot signaling games, without use of equilibrium refinement criteria specifically designed to pick out the desired equilibria.

Here we propose exactly such a derivation, by revisiting the assumption in our base model regarding the nature of literal meaning in the absence of prior conventions. Earlier in this section we assumed that the absence of prior convention should be represented as a single lexicon $\mathcal{L}$ in which all utterances have the all-*true* (tautological) meaning. We revise that assumption in two respects. First, for any given utterance we allow $\mathcal{L}_u$ to assume either truth value $\{0,1\}$ for each meaning, allowing the lexicon to assign non-trivial semantic content to utterances. Second, we allow for *lexical uncertainty*, where the speaker and listener can reason about distributions over multiple lexica. In the signaling game described above, for example, lexicon $\mathcal{L}^1$ might assign the meaning *likely* to "expensive" and the all-*true* meaning to "cheap", whereas lexicon $\mathcal{L}^2$ might assign the meaning *unlikely* to "expensive" and the same all-*true* meaning to "cheap". The absence of prior conventions then means that the *marginal* interpretation, across lexica, is the same for all utterances.

Including lexical uncertainty generalizes the previous model; the base listener L_0 remains unchanged from equation 1, but the more sophisticated speaker and listener are defined by:

$$S_n(u|m,\mathcal{L}) \propto e^{\lambda U_n(u|m,\mathcal{L})} \tag{6}$$

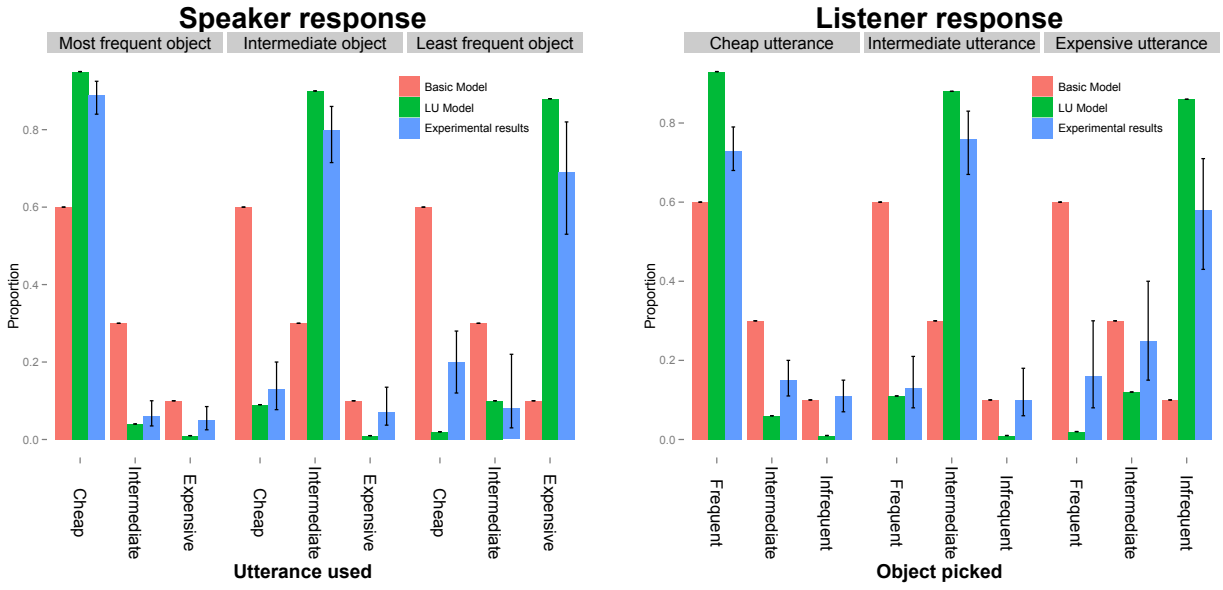
$$L_n(m|u) \propto \sum_{\mathcal{L}} P(m)P(\mathcal{L})S_{n-1}(u|m,\mathcal{L}) \tag{7}$$

where

$$U_n(u|m,\mathcal{L}) = \begin{cases} \log(L_0(m|u,\mathcal{L})) - c(u) & \text{if } n=1 \\ \log(L_{n-1}(m|u)) - c(u) & \text{if } n>1. \end{cases} \tag{8}$$

We take $P(\mathcal{L})$ to be the uniform distribution over all seven logically possible lexica in which every utterance assigns "true" to at least one meaning and every meaning is assigned "true" by at least one utterance.

Because this new *lexical-uncertainty* model reduces to the base model when conventions for literal meanings are already established and there is only a single lexicon $\mathcal{L}$, the new model continues to properly handle specificity implicature (Figure 2, green bars). Furthermore, the new model derives Horn's principle. Consider the case we proposed above with two lexica, $\mathcal{L}^1$ interpreting "expensive" as *likely* and $\mathcal{L}^2$ interpreting "expensive" as *unlikely* (and both giving a trivial interpretation to "cheap"). Due to the role of the prior $P(m)$, the base listener L_0 associates "cheap" with *likely* for both lexica. Now consider two speakers who can use the expensive utterance to precisely communicate their meaning: speaker $S_1(\cdot|likely,\mathcal{L}^1)$ who wants to communicate *likely* and is using



(a) (b)

Figure 3: Experiment 2 results and model predictions. (a) Speaker's message choice given that she needed to convey the most frequent (left), intermediate (center), or least frequent object (right). The y-axis is the proportion of trials on which each message was chosen. Error bars are 95% confidence intervals. (b) Listener's object choice after receiving the cheapest (left), intermediate (center), or most expensive message (right). The y-axis is the proportion of trials on which each object was chosen.

lexicon $\mathcal{L}^1$ and speaker $S_1(\cdot|unlikely, \mathcal{L}^2)$ who wants to communicate *unlikely* and is using lexicon $\mathcal{L}^2$. For the speaker $S_1(\cdot|likely, \mathcal{L}^1)$, the extra precision of "expensive" is not valuable, because the base listener L_0 will also interpret "cheap" as *likely*. However, for the speaker $S_1(\cdot|unlikely, \mathcal{L}^2)$, the extra precision of "expensive" *is* valuable, because it overrides the base listener's prior bias against *unlikely*. This breaks the symmetry and leads L_2 to start to prefer the efficient mapping, a preference that gets magnified by S_3's reasoning and continues to get magnified at higher levels of recursive inference.

The lexical uncertainty model can predict correct Horn equilibria beyond the case with two meanings and utterances; predictions (approximated by averaging over a finely gridded approximation to the parameter space) for the case with three meanings and utterances are shown in Figure 3.

Experiment 2

In Experiment 2 we investigated whether people can coordinate on Horn's principle in the absence of prior linguistic conventions. This experiment was designed to be a minimal test of this question: people were placed in the simplest setting in which Horn's principle is possible. People played a communication game with a partner as in Experiment 1. There were three possible meanings for the speaker to communicate, which differed in how frequently they appeared in the game. The speaker was able to communicate the intended meaning by sending one of three messages, which differed only in their cost to the speaker—none was iconic. If the same reasoning that gives rise to Horn's principle extends to this novel setting, then we expect the speaker and listener to coordinate on the efficient mapping of meanings to messages.

Methods:

We recruited 140 participants from Amazon Mechanical Turk, who were paid for participation in addition to bonus payments described below.

The interface and instructions for the experiment were very similar to Experiment 1. Participants were told that they landed on an alien planet containing three kinds of objects and three messages that could be used to communicate these objects. They were told that these objects occurred with different frequencies on the planet; one occurred 60% of the time, one occurred 30%, and the last occurred 10%. Of the three messages, one was free, one cost \$0.01, and the last cost \$0.02. The object frequencies and message costs were randomized between subjects. Participants received 10 practice rounds without a partner to familiarize them with the interface, object frequencies, and message costs, but received no feedback during these rounds.

Each participant played 5 rounds of the game with a partner randomly assigned each round. The game was the same as in Experiment 1, with two changes. The object that the speaker needed to communicate was randomly sampled according to the frequency of the objects on the alien planet (so, e.g., the most frequent object was sampled 60% of the time). Second, the speaker was charged the cost of the message sent.

Table 1: Experiment 2 analyses

Role	Response	Comparison	t-value	p-value
Listener	Frequent object	Cheap utterance > intermediate, expensive	6.07	0.001
	Intermediate object	Intermediate utterance > cheap, expensive	5.31	0.001
	Unlikely object	Expensive utterance > cheap, intermediate	5.55	0.001
Speaker	Cheap utterance	Frequent object > intermediate, unlikely	8.27	0.001
	Intermediate utterance	Intermediate object > frequent, unlikely	6.21	0.001
	Expensive utterance	Unlikely object > frequent, intermediate	5.55	0.001

Results and Discussion

Human speaker and listener choices are shown in figure 3, alongside the predictions of our base model (pink bars), which does not predict the Horn equilibrium, and our lexical-uncertainty model (green bars), which does. We first analyzed whether listeners interpreted messages according to the efficient mapping, by carrying out three mixed logit regressions with random intercepts for participants. We analyzed whether, e.g., listeners responded with the frequent object more often when they received the cheap utterance than the other utterances. These comparisons are shown in Table 1. They show that the listener's responses were consistent with the efficient mapping. We next analyzed whether speakers chose messages efficiently. We addressed this question in a similar manner to the previous one, carrying out three mixed logit regressions with random intercepts for participants. The comparisons in table 1 show that the speaker was more likely to use the cheap utterance given the frequent object than given the other objects, and similarly for the other utterances.

By design, each participant only played five rounds of the game. This was done to ensure that observed efficiency effects were due to cooperative reasoning, and not due to language evolution. To validate this design, we analyzed whether participants played differently on the first round than on future rounds. To do this, we performed by-subject ANOVAs on the speaker and listener responses to determine whether there were main effects or interactions from the first round. For five of the six speaker and listener response types, there was no main effect of the first round or interaction with the object to communicate or message received ($p>0.05$). For the intermediate-cost message, there was a small but significant interaction between the first round and the object to communicate ($p<0.05$). These analyses provide evidence that learning or language evolution are not driving our results.

These results provide evidence that people can construct efficient strategies for communication on-line, in the absence of prior linguistic conventions. This suggests that Horn's principle, as it applies to ordinary language use, may arise from cooperative reasoning between people trying to communicate with each other, rather than from language evolution.

Discussion

We have investigated two kinds of pragmatic contrast effects. First, we looked at the effect of varying the specificity, or informativeness, of the utterances available to the speaker. A simple model of social cognition was able to account for the strengthening of a non-specific utterance's interpretation. We found in Experiment 1 that listeners inferred this strengthened interpretation, and that speakers anticipated this, in the absence of any prior linguistic conventions.

We next turned to Horn's principle and the effect of varying the relative cost or complexity of utterances. We first found that the simple model of social cognition cannot explain Horn's principle, because it does not have the resources to exploit the asymmetry between more and less costly utterances. However, once the model included uncertainty about the underlying lexicon, it was able to predict Horn's principle, and explain this asymmetry: under the model, only someone who wanted to communicate an unlikely meaning would have an incentive to use an expensive utterance. Notably, this is the first proposed model of Horn's principle which does not rely on specialized equilibrium selection criteria. In Experiment 2, we found evidence that people expect costlier utterances to correspond to less frequent meanings, as predicted by Horn's principle.

Our results suggest that in two important cases, people's pragmatic knowledge extends beyond learned grammatical knowledge to novel, non-linguistic communicative scenarios. While we have not settled the question of how people arrive at ordinary pragmatic inferences, our experimental and modeling results do provide evidence that linguistic conventions are not necessary for them—social cognition will suffice.

References

Chen, Y., Kartik, N., & Sobel, J. (2008). Selecting cheap-talk equilibria. *Econometrica*, *76*(1), 117–136.

Chierchia, G. (2004). Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. *Structures and beyond*, *3*, 39–103.

Cho, I., & Kreps, D. (1987). Signaling games and stable equilibria. *The Quarterly Journal of Economics*, *102*(2).

Clark, H. (1996). *Using language.* Cambridge University Press (Cambridge England and New York).

Frank, M. C., Goodman, N. D., Lai, P., & Tenenbaum, J. B. (2009). Informative communication in word production and word learning. In *Proc. cog. sci. soc.*

Franke, M. (2009). Interpretation of optimal signals. *New perspectives on games and interaction*, 297–310.

Grice, H. (1975). Logic and conversation. *1975*, 41–58.

Horn, L. R. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (Ed.), *Meaning, form, and use in context: Linguistic applications* (pp. 11–42).

Jäger, G., & Ebert, C. (2009). Pragmatic rationalizability. In *Proceedings of sinn und bedeutung* (Vol. 13, pp. 1–15).

Lewis, D. (1969). *Convention: a philosophical study*. Harvard University Press.

Sutton, R., & Barto, A. (1998). *Reinforcement learning: An introduction* (Vol. 28). Cambridge Univ Press.

van Rooij, R. (2004). Signalling games select horn strategies. *Linguistics and Philosophy*, *27*(4), 493–527.

van Rooij, R. (2008). Games and quantity implicatures. *Journal of Economic Methodology*, *15*(3), 261–274.