Predicting pragmatic reasoning in language games

Michael C. Frank[1*] & Noah D. Goodman[1]

[1]Department of Psychology, Stanford University

[*]To whom correspondence should be addressed; E-mail: mcfrank@stanford.edu.

Online abstract. **One of the most astonishing features of human language is its capacity to convey information efficiently in context. Many theories provide informal accounts of communicative inference, yet there have been few successes in making precise, quantitative predictions about pragmatic reasoning. We examine judgments about simple referential communication games, modeling behavior in these games by assuming that speakers attempt to be informative, and that listeners use Bayesian inference to recover speakers' intended referents. Our model provides a close, parameter-free fit to human judgments, suggesting that using information-theoretic tools to predict pragmatic reasoning may lead to more effective formal models of communication.**

One of the most astonishing features of human language is its ability to convey information efficiently in context. Each utterance need not carry every detail; instead, listeners can infer speakers' intended meanings by assuming utterances convey only relevant information. These communicative inferences rely on the shared assumption that speakers are informative, but not more so than is necessary given the communicators' common knowledge and the task at hand. Many theories provide high-level accounts of these kinds of inferences (*1–3*), yet—perhaps be-

cause of the difficulty of formalizing notions like "informativeness" or "common knowledge"—there have been few successes in making quantitative predictions about pragmatic inference in context.

We address this issue by studying simple referential communication games, like those described by Wittgenstein (*4*). Participants see a set of objects and are asked to bet which one is being referred to by a particular word. We model human behavior by assuming that a listener can use Bayesian inference to recover a speaker's intended referent r_S in context C, given that the speaker uttered word w:

$$P(r_S|w, C) = \frac{P(w|r_S, C)P(r_S)}{\sum_{r' \in C} P(w|r', C)P(r')}. \quad (1)$$

This expression is the product of three terms: the prior probability $P(r_S)$ that an object would be referred to; the likelihood $P(w|r_S, C)$ that the speaker would utter a particular word to refer to the object; and the normalizing constant, a sum of these terms computed for all referents in the context.

We define the prior probability of referring to an object as its contextual salience. This term picks out not just perceptually- but also socially- and conversationally-salient objects, capturing the common knowledge that speaker and listener share, as it affects the communication game. Because there is no *a priori* method for computing this sort of salience, we instead measure it empirically (*5*).

The likelihood term in our model is defined by the assumption that speakers choose words to be informative in context. We quantify the informativeness of a word by its surprisal—an information-theoretic measure of how much it reduces uncertainty about the referent. By assuming a rational actor model of the speaker, with utility defined in terms of surprisal, we can derive the regularity that speakers should choose words proportional to their specificity (*6, 7*):

$$P(w|r_S, C) = \frac{|w|^{-1}}{\sum_{w' \in W} |w'|^{-1}}, \tag{2}$$

where $|w|$ indicates the number of objects to which word w could apply and W indicates the set of words that apply to the speaker's intended referent.

In our experiment, three groups of participants each saw communicative contexts consisting of sets of objects varying on two dimensions (Fig. 1A). We systematically varied the distribution of features on these dimensions. To minimize the effects of particular configurations or features, we randomized all other aspects of the objects for each participant. The first group (*speaker*) bet on which word a speaker would use to describe a particular object, testing the likelihood portion of our model. The second group (*salience*) was told that a speaker had used an unknown word to refer to one of the objects, and asked to bet which object was being talked about, providing an empirical measure of the prior in our model. The third group (*listener*) was told that a speaker had used a single word (e.g. "blue") and again asked to bet on objects, testing the posterior predictions of our model.

Mean bets in the speaker condition were highly correlated with our model's predictions for informative speakers ($r = .98$, $p < .001$; Fig. 1B, open circles). Judgments in the salience and listener conditions were not themselves correlated with one another ($r = .19$, $p = .40$), but when salience and informativeness terms were combined via our model, the result was highly correlated with listener judgments ($r = .99$, $p < .0001$, Fig. 1B, closed circles). This correlation remained highly significant when predictions of 0 and 100 were removed ($r = .87$, $p < .0001$). Fig. 1C shows model calculations for one arrangement of objects.

Our simple model synthesizes and extends work on human communication from a number of different traditions, including early disambiguation models (*8*), game-theoretic signaling models (*9*), and systems for generating referring expressions (*10*). The combination of

an information-theoretic definition of "informativeness" along with empirical measurements of common knowledge enables us to capture some of the richness of human pragmatic inference in context.

References

1. H. Grice, in *Syntax and Semantics, vol. 3*, P. Cole and J. Morgan (Academic Press, New York, NY, 1975), 41–58.

2. D. Sperber, D. Wilson, *Relevance: Communication and Cognition* (Harvard University Press, Cambridge, MA, 1986).

3. H. Clark, *Using Language* (Cambridge University Press, Cambridge, UK, 1996).

4. L. Wittgenstein, *Philosophical Investigations* (Blackwell Publishers, Oxford, UK, 1953).

5. H. Clark, R. Schreuder, S. Buttrick, *J. Verb. Learn. Verb. Behav.* **22**, 245–258 (1983).

6. Materials and methods are available as supplementary material on *Science* Online.

7. F. Xu, J. Tenenbaum, *Psychol. Rev.* **114**, 245–272 (2007).

8. S. Rosenberg, B. Cohen, *Science* **145**, 1201–1203 (1964).

9. A. Benz, G. Jäger, R. Van Rooij, Eds., *Game theory and pragmatics* (Palgrave Macmillan, Hampshire, UK, 2005).

10. R. Dale, E. Reiter, *Cognit. Sci.* **19**, 233–263 (1995).

Supplementary Materials

www.sciencemag.org

Materials and Methods

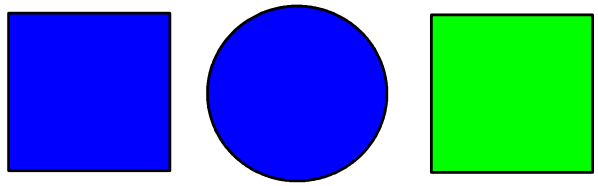
Supplementary Text

Figure Legend

(A) An example stimulus from our experiment, with instructions for speaker, listener, and salience conditions. (B) Human bets on the probability of a choosing a term (speaker condition, $N = 206$) or referring to an object (listener condition, $N = 263$), plotted by model predictions. Points represent mean bets for particular terms and objects for each context type. The red line shows the best linear fit to all data. (C) An example calculation in our model, for the context type shown in panel A. Empirical data from the salience condition constitutes the prior term, $N = 20$ (top); this is multiplied by the model-derived likelihood term (middle). The resulting posterior model predictions (normalization step not shown) are plotted alongside human data from the listener condition, $N = 24$ (bottom). All error bars show 95% confidence intervals.

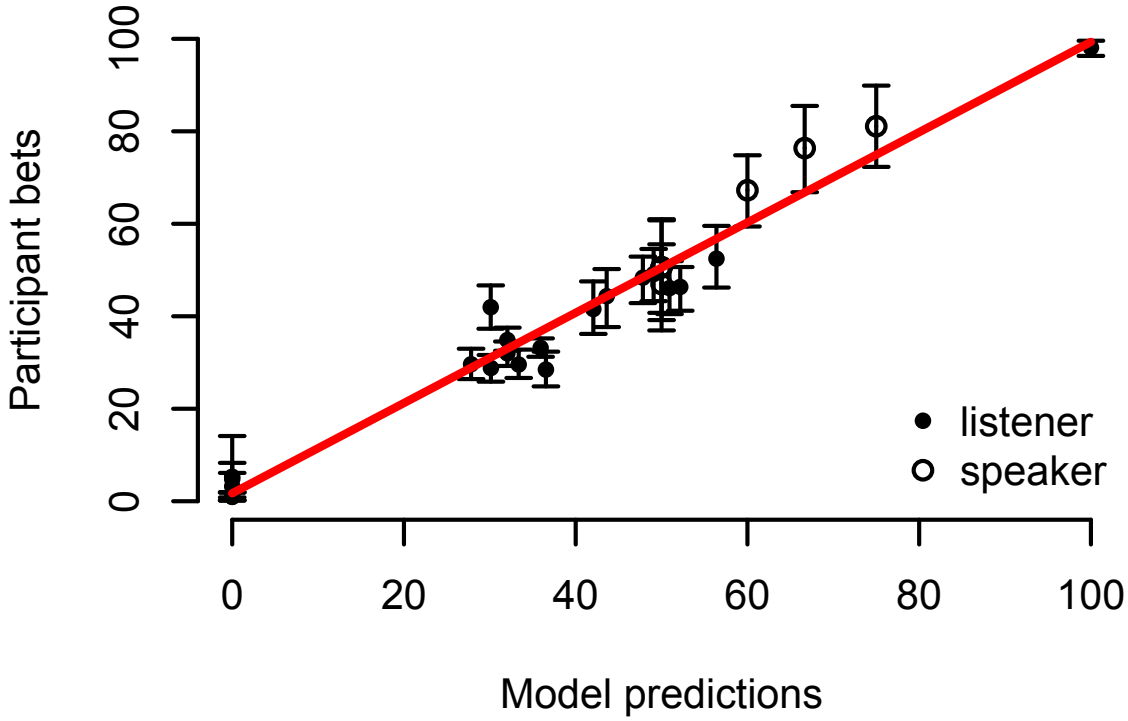
A

Speaker: Imagine you are talking to someone and you want to refer to the middle object. Which word would you use, "**blue**" or "**circle**"?

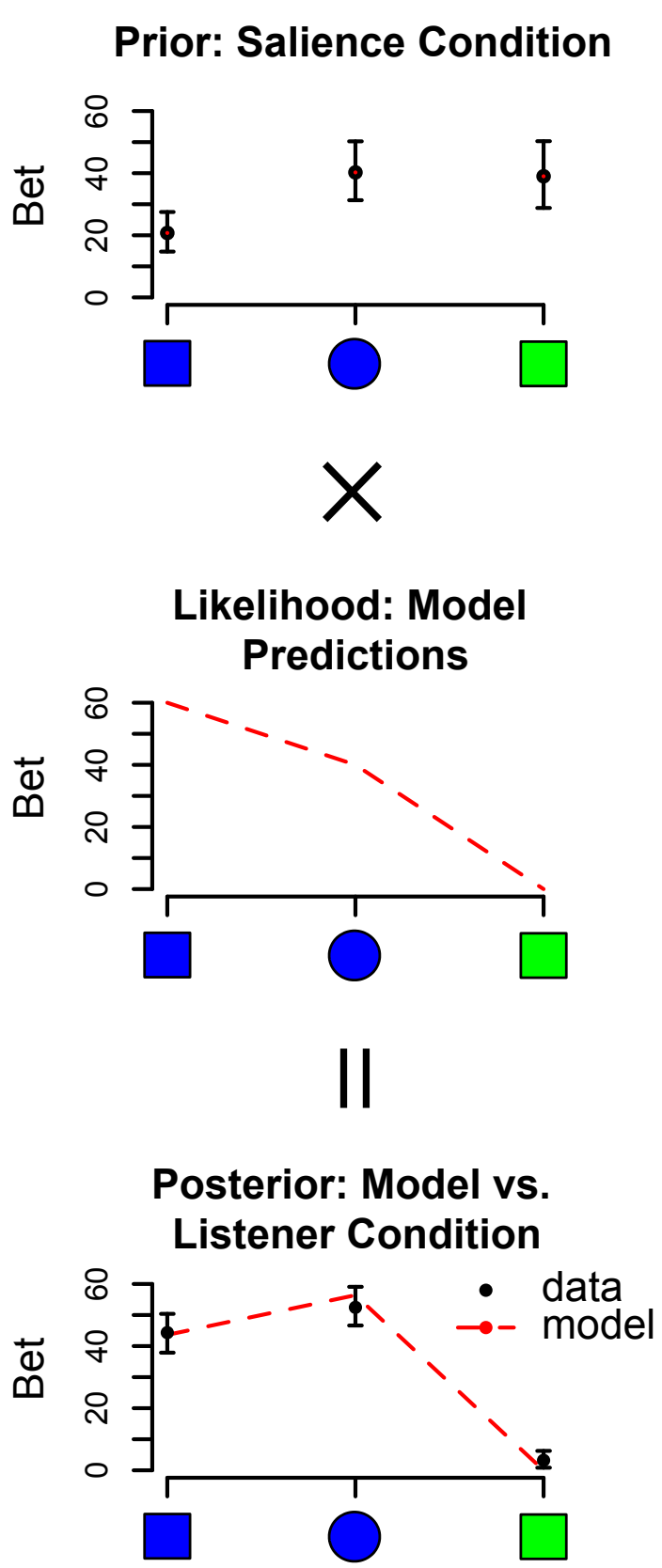


Listener/Salience: Imagine someone is talking to you and uses [the word "**blue**"/a word you don't know] to refer to one of these objects. Which object are they talking about?

B



C



Supplementary Materials for

Predicting Pragmatic Reasoning in Language Games

Michael C. Frank and Noah D. Goodman

correspondence to: mcfrank@stanford.edu

This PDF file includes:

Materials and Methods

Participants were 745 individuals in the United States, recruited via Amazon's Mechanical Turk (`www.mturk.com`, an online crowd-sourcing tool): 206 in the speaker condition, 276 in the salience condition and 263 in the listener condition. We posted a total of 900 individual trials. Each trial in the final sample for each condition was from a unique participant who passed a manipulation check and whose bets added to 100. The manipulation check asked participants to report the number of objects with each of the features of the target in the dimension of interest (e.g., "How many objects are blue?"). If participants contributed more than one trial, only their first was included and their additional trials were reposted for other workers to complete.

Each Mechanical Turk HIT consisted of a single web page displaying a randomly-generated set of three objects, and each object was assigned a color (red / blue / green), a shape (circle / square / cloud), and a texture (solid / polka-dot / striped) feature. In each object set, two of these feature dimensions were chosen to vary, while the third was held constant. The critical manipulation was the distribution of feature values on these two dimensions: we systematically varied whether one, two, or all three of the objects shared values with a target object.

This manipulation resulted in seven distinct context types; all others reduce to these context types by symmetry. We notate these context types as: # of objects with same value of feature 1 as target / # of objects with same value of feature 2 as target. For example, 1/3 indicates that the target object was the only object in the set with a particular value on feature 1, but all three objects shared the target's value on feature 2. This manipulation resulted in seven distinct conditions: 1/1, 1/2, 1/3, 2/2 with both features overlapping on two objects, 2/2 with both features overlapping on one object (pictured in Fig. 1C), 2/3, and 3/3. Fifty random trials were generated for each of the 6 unique numerical conditions for each condition (speaker, salience, and listener), and then the two 2/2 conditions were separated in the analysis because our model generated distinct predictions for them.

In the speaker condition, the target object was indicated via a dotted line around it (since target position was always randomized).

Supplementary Text

Model Derivation

Speaker and listener are interacting in a shared referential context using a shared vocabulary. The context consists of a set of objects, $C = \{o_1 ... o_n\}$. Each word in the vocabulary, $V = \{w_1 ... w_m\}$, has a meaning (also shared by speaker and listener), which is a Boolean function on objects. The speaker acts rationally according to Bayesian decision theory by choosing words in proportion to their expected utility:

$$P(w|r_s, C) \propto e^{\alpha U(w;r_S,C)}. \tag{S1}$$

The decision noise parameter α measures the speaker's deviation from optimal action selection. We set $\alpha = 1$ to recover a standard Luce choice rule.

The speaker's goal is to choose the utterance that is both maximally informative with respect to the speaker's intended referent and also maximally inexpensive to speak, so utility is defined as

$$U(w; r_S, C) = I(w; r_S, C) - D(w) \tag{S2}$$

where $I(w; r_S, C)$ represents the informativeness of an utterance with respect to the speaker's intended referent and $D(w)$ represents its cost. We assume $D(w)$ is constant, since all words in our stimuli are randomized (and roughly matched for complexity), but note that in other situations, cost may be affected by word length, utterance length, frequency, and other factors known to play a role in speech production. Future work should examine the role of this cost factor in interpretive inferences.

We quantify the informativeness of a word using the self-information, or surprisal: $I_p(x) = -\log(p(x))$, which measures of how much information is gained by observing a particular sample x from a known distribution $p(x)$. Speaker's utility decreases with surprisal: $I(w; r_S, C) = -I_{\tilde{w}_C}(r_S)$, where $\tilde{w}_C$ is the distribution over objects that would come from a literal interpretation of w in context C.

If listeners interpret the utterance w literally, assigning zero probability to objects for which the word is false, they assign equal probability to each object consistent with w. This distribution over objects can be written:

$$\tilde{w}_C(o) = \begin{cases} \frac{1}{|w|} & \text{if } w(o) = \texttt{true} \\ 0 & \text{otherwise.} \end{cases} \tag{S3}$$

Therefore, by Equations S1 – S3, we have

$$P(w|r_S, C) = \frac{e^{-(-\log(|w|^{-1}))}}{\sum_{w' \in V \text{ st. } w'(r_s)=\texttt{true}} e^{-(-\log(|w'|^{-1}))}}, \tag{S4}$$

which reduces to Equation 2 in the main text, also known as the "size principle" (*7*). Thus, in our experiments, the speaker's abstract goal of being informative reduces to a simple formulation: choose a word that applies to the referent and picks out a relatively smaller section of the context. Listeners may then use this model of a speaker as their likelihood function, to be combined with prior information about contextual salience as in Equation 1 in the main text.