www.sciencemag.org/cgi/content/full/336/6084/998/DC1

Supplementary Materials for

Predicting Pragmatic Reasoning in Language Games

Michael C. Frank* and Noah D. Goodman

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

Published 25 May 2012, *Science* **336**, 998 (2012)
DOI: 10.1126/science.1218633

This PDF file includes:

Materials and Methods
Supplementary Text

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

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

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.