Informative Communication in Word Production and Word Learning

Michael C. Frank, Noah D. Goodman, Peter Lai, and Joshua B. Tenenbaum
{mcfrank, ndg, peterlai, jbt}@mit.edu
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

Abstract

Language does not directly code facts about the world. Instead, speakers and listeners rely on shared assumptions to allow them to communicate more efficiently. Writers like Grice and Sperber & Wilson have proposed that communication is assumed to be "informative" or "relevant," but the predictions of these accounts are often informal or post-hoc. Here we propose a formal analogue to these accounts: that communicators choose what they want to say by how informative it would be about their intended meaning. We derive quantitative predictions about how this assumption would be used in language production and learning and test these predictions via two experiments. This work takes a first step towards formalizing the pragmatic assumptions necessary for effective communication in under-constrained, real-world situations.

Keywords: Language acquisition; Bayesian modeling; Communication

Introduction

How does language work to communicate information from one person to another? Perhaps language is simply a code for facts about the world. On this kind of *coding* view of communication, all the information necessary to understand an utterance is contained within it. Speakers utter linguistic expressions equivalent to their intended meanings and listeners simply decode these expressions to recover their content. There are a profusion of examples of language use, however, which can be natural and easy to understand but are not easily explained by a naive coding model:

(1) The statement "I ate some of the cookies." (Intended meaning: I ate some and not all of the cookies).

(2) The declaration "No." (Intended meaning: I can tell you want to pinch him, but don't do it).

(3) The contextual introduction of a new word "Can I have the glorzit?" (Intended meaning: pass me that thing, which happens to be called a "glorzit").

Philosophers and linguists interested in this problem have suggested that language relies on shared assumptions about the nature of the communicative task. Grice (1975) proposed that speakers follow (and are assumed by comprehenders to follow) a set of maxims, such as "be relevant", or "make your contribution to the conversation as informative as necessary." Sperber & Wilson (1986) have suggested that there is a shared "Principle of Relevance" which underlies communication. Clark (1996) has argued that communication proceeds by reference to a shared "common ground."

Though these proposals differ in their details, they share a basic assumption that communicators are not simply coding and decoding meanings. Instead, listeners are making inferences about speakers' intentions, taking into account the words they utter and the context of their utterances. This kind of *intentional inference* framework for language seems much more promising for explaining phenomena like (1-3). But although these ideas seem intuitively correct, the difficulty of formalizing notions like "relevance" has largely kept them from making contact with computational theories of language use and acquisition.

The goal of this paper is to begin to address this issue by proposing a computational framework for intentional inference. This framework relies on a shared assumption that communications are informative given the context. Although the basis of our framework is general, making predictions within it requires a model of the space of possible meanings and how they map to natural language expressions. Thus, in order to make a first test of our framework, we study simple games that are similar to the "language games" proposed by Wittgenstein (1953).

In the language games we study, the shared task of communicators is to identify an object from a set using one or a few words. This very restricted task allows us to define the possible meanings that communicators entertain. We then use our framework to make predictions about the meaning and use of single words. This move allows us to define an intuitive mapping between words and meanings: that a word stands for the subset of the context it picks out (its extension). Although these two simplifications do bring our tasks further away from natural language use, they also allow us to derive strong quantitative predictions from our framework.

The outline of the paper is as follows. We first use our framework to derive predictions for speakers and language learners who assume informative communication in an inferential framework. We then test our framework as an account of two different kinds of tasks. Experiment 1 examines, in a simple survey task, whether learners who are inferring the meaning of a novel word assume that speakers are being informative in choosing the word they produce. Experiment 2 tests whether, in a more naturalistic production task, speakers' word choice is in fact related to the informativeness of the word they pick.

Modeling Informative Communication

Consider the context in Figure 1, representing the context in a language game. Imagine an English speaker in this game who is told to use a single word to point out the red circle.

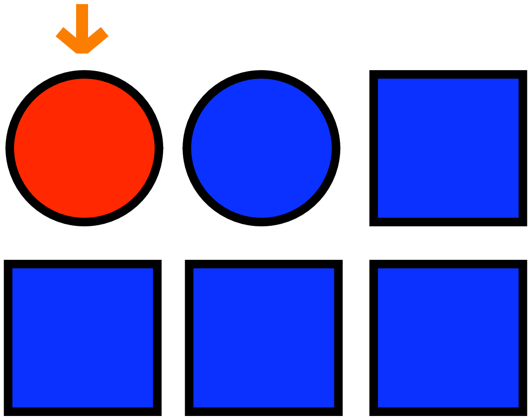


Figure 1: An example context: one red circle, one blue circle, and four blue squares. The arrow shows the object the speaker intends to talk about.

Intuitively, she is likely to use the word "red," since it would not be clear which object she was talking about if she used the word "circle." Working from this intuition, a language learner—who knew that the speaker was pointing out the red circle (perhaps because of some non-linguistic marker of intention, like a point or an eye-movement)—could make a very informed guess about the meaning of the word "red."

The intuition that a speaker would be likely to use the word "red" to talk about the red circle seems to come from an assumption that, in the language of Grice (1975), speakers are choosing their words helpfully, in order to best inform listeners of their intentions. If speakers did not try to communicate informatively, they would not necessarily choose labels that distinguished their intended referent from other possible objects. In our game, an uninformative speaker (who still respected the truth conditions of their language) might just as well have chosen to talk about the shape of the red circle as its color; correspondingly, a learner who did not assume an informative speaker would not be able to infer what the word "red" meant.

In the following sections, we formalize these intuitions through an inferential model of language within this restricted world. We model the speaker as selecting speech acts in order to be informative, and derive predictions both for speakers and for learners who assume this about speakers.[1]

The Informative Communication Framework

We assume that there is a fixed context C, consisting of some set of objects $o_1 ... o_m$, and that possible meanings are probability distributions over C—that is, a meaning assigns a probability over each object in C. In the game described above, meanings simply carry information about which object is the intended referent, though in a more complex task they might be distributions over propositions, for example (Piantadosi et al., 2008). We use this space of possible meanings for both the intended meaning of the speaker (which will be a delta distribution when the referent is known), and the meanings of words. By using distributional meanings, we are able to use notions of informativeness from information theory in formulating the communicative goals of the speaker.

Imagine that there is a vocabulary $V = \{w_1, ..., w_p\}$, and each word has a truth-functional meaning: a Boolean function over objects, indicating whether the word applies to that object. The extension of word w in context C is the set of objects $\{o \in C | w(o) = 1\}$ that the word applies to; denote by $|w|$ the size of a word's extension. We define the meaning of w in context C to be the distribution:

$$\tilde{w}_C(o) = \begin{cases} \frac{1}{|w|} & \text{if } w(o) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Rational speaker We assume the speaker acts rationally, according to Bayesian decision theory: she chooses a speech act (word choice) in order to (soft-)maximize utility:

$$P(w|M_S, C) \propto e^{\alpha U(w; M_S, C)} \tag{5}$$

where M_S is her intended meaning and U is a utility function. For us, intended meanings will be delta distributions picking out her intended referent precisely, but the model will naturally handle vague intended meanings (which might arise, for instance, from incomplete knowledge). (The decision noise parameter α measures the speakers deviation from optimal. For all computations in the current paper we set $\alpha = 1$, which recovers the standard Luce choice rule.) The speaker's goal is to choose the word which is informative about M_S.

The Kullback-Leibler divergence between two distributions X and Y, written $D_{KL}(X||Y)$, is the expected amount of information about X that is not contained in Y (Cover & Thomas, 2006). We formalize the goal of "informativeness" by assuming that the speaker's utility increases as the KL divergence between M_S and the literal meaning of the chosen word decreases:

$$U(w; M_S, C) = -D_{KL}(M_S || \tilde{w}_C) + F \tag{6}$$

where F represents other factors (such as utterance complexity) which affect the speaker's utility. Assuming for the time being that $F = 0$:

$$P(w|M_S, C) = \frac{e^{-\alpha D_{KL}(M_S || \tilde{w}_C)}}{\sum\limits_{w' \in V} e^{-\alpha D_{KL}(M_S || \tilde{w}'_C)}} \tag{7}$$

Equation 7 simplifies greatly in simple language games like the one pictured in Figure 1. The speaker's intended meaning is a single object o_S (the value of M_S is 1 for o_S, 0 for all other objects). Thus:

$$\begin{aligned} D_{KL}(M_S || \tilde{w}_C) &= \sum_{o \in C} M_S(o) \log \frac{M_S(o)}{\tilde{w}(o)} \\ &= -\log(\tilde{w}_C(o_S)) \end{aligned} \tag{8}$$

[1] One case which we do not treat here is the case of a teacher who is searching for the *best* example of a word to show a learner. This case is discussed in detail in Shafto & Goodman (2008), and we believe the current framework is compatible with their analysis.

Substituting Equation 8 into Equation 7 gives:

$$P(w|M_S,C) = \frac{\tilde{w}_C(o_S)^{\alpha}}{\sum\limits_{w' \in V} \tilde{w}'_C(o_S)^{\alpha}} \tag{9}$$

By Equation 4:

$$P(w|M_S,C) \propto \begin{cases} |w|^{-\alpha} & \text{if } w(o) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Thus, given the set of simplifying assumptions we have made, the very abstract goal of "being informative" reduces to a simple formulation: choose words which pick out relatively smaller sections of the context. This recovers the "size principle" of Tenenbaum & Griffiths (2002) and Xu & Tenenbaum (2007).[2]

Ostensive word learning We next show how a Bayesian learner equipped with the theory of helpful speakers captured in Equation 10 can leverage this knowledge to learn words in an unknown language. A real language learner often has uncertainty about both the speaker's meaning, M_S, and the lexicon, L, the mappings between words and meanings (Frank et al., in press). Although our framework can be extended to this case, we focus here on a simpler case that is close to ostensive learning: the learner knows M_S and has only to infer facts about L. In this case, the learner knows which object the speaker means to talk about, and can use the assumptions of informative communication to infer the meaning of the spoken word (i.e. which feature of the object it refers to). By Bayes' rule:

$$P(L|w,M_S,C) \propto P(w|L,M_S,C)P(L) \tag{11}$$

For simplicity let us assume that the object has two features f_1 and f_2, that there are two words in the language w_1 and w_2, and that there are only two possible lexicons $L_1 = \{w_1{=}f_1, w_2{=}f_2\}$ and $L_2 = \{w_1{=}f_2, w_2{=}f_1\}$. Further, assume a uniform prior on vocabularies. Then:

$$\begin{aligned} P(L_1|w_1,M_S,C) &= \frac{P(w_1|L_1,M_S,C)}{P(w_1|L_1,M_S,C)+P(w_1|L_2,M_S,C)} \\ &= \frac{|f_1|^{-1}}{|f_1|^{-1}+|f_2|^{-1}} \end{aligned} \tag{12}$$

Experiment 1

In order to make a first test of our framework, we used an experimental paradigm based on the "red circle" example above. We created a web survey which asked participants to imagine encountering a display like Figure 1 and seeing a speaker of a foreign language indicate one of the objects and say a word in her language. We then asked asked the participants to make judgments about the meaning of that word. In order to elicit a continuous, quantitative judgment, we asked participants to "bet," splitting $100 between the two possible meanings. This betting measure gives us an estimate of speakers' subjective probability.

We then attempted to predict participants' mean bets across a range of different contexts. For example, in Figure 1, imagine that the speaker points to the red circle and says "lipfy" (a novel word w that you have never heard before). We used Equation 12 to calculate the probability that learners judge that w means `red` as opposed to `circular`:

$$\begin{aligned} P(w = f_1|M_S,C) &= \frac{|\texttt{red}|^{-1}}{|\texttt{red}|^{-1}+|\texttt{circular}|^{-1}} \\ &= \frac{\frac{1}{1}}{\frac{1}{1}+\frac{1}{2}} = \frac{2}{3} \end{aligned}$$

Thus, our prediction is that learners should bet around $67 that "lipfy" means red.

Materials and Methods

Participants Seven hundred participants responded to solicitations through MIT-related email lists and were compensated via entry in a drawing to win an iPod Shuffle.

Procedure Participants navigated to a webpage on which they were given a short backstory about the experiment. They were told to imagine that they were visiting a foreign country and that the display they were shown was a display in a market. They were told that one of the items (indicated by a square around it) was being described by the merchant so as to teach them a word, and that their task was to guess the meaning of the word that the merchant used.

In each of three trials on the page, six simple objects were shown. The objects in each trial varied randomly on two binary-valued dimensions picked from a larger set of features (red/blue, big/small, striped/polka-dot, circular/square) and whose other properties were constant on these dimensions. For example, in Figure 1, size and texture are fixed but objects varied on color and shape. All trials were constructed such that different properties were used for each trial and participants were not able to make mutual-exclusivity judgments between trials.

On each trial, participants were told that the speaker was talking about one particular object using a novel word (e.g. "lipfy") and asked to split a bet of $100 dollars between the two attributes that "lipfy" could refer to (e.g. `red` or `circular`). Different novel words were used in each trial.

Results & Model Fits

Results are plotted in Figure 2. Since we had randomized the dimensions used in each trial, we averaged across this aspect of the data and considered only the distribution of bets

[2] This principle was originally derived by Shepard (1987) as a description of appropriate generalization behavior within psychological spaces. Our work here can be thought of as an alternate derivation of the size principle—based on premises about the communicative task, rather than about the structure of generalization—that licenses its application to the kinds of cases that we have treated here.

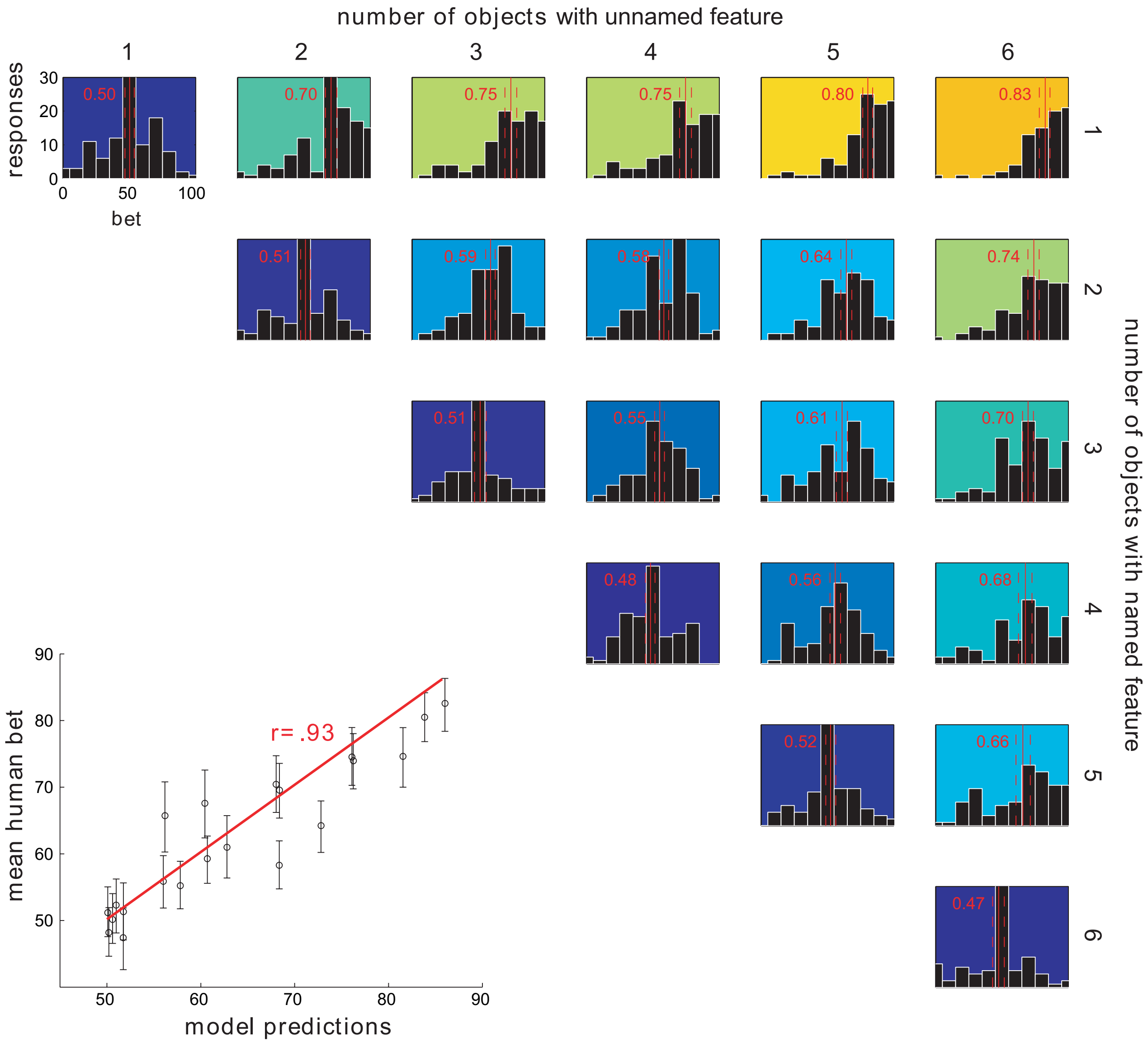


Figure 2: Each subplot shows the histogram of participants' bets, along with the mean and 95% confidence intervals shown in red. Confidence intervals are computed via non-parametric bootstrap. Plots are arranged by the number of objects with the named and unnamed features (e.g., in Figure 1, "red" and "circular"). Plot background colors are proportional to the mean bet. The inscribed plot (lower-left) shows mean bets plotted by model predictions with 95% confidence intervals. X positions are jittered slightly to avoid overplotting. Line of best fit is shown in red.

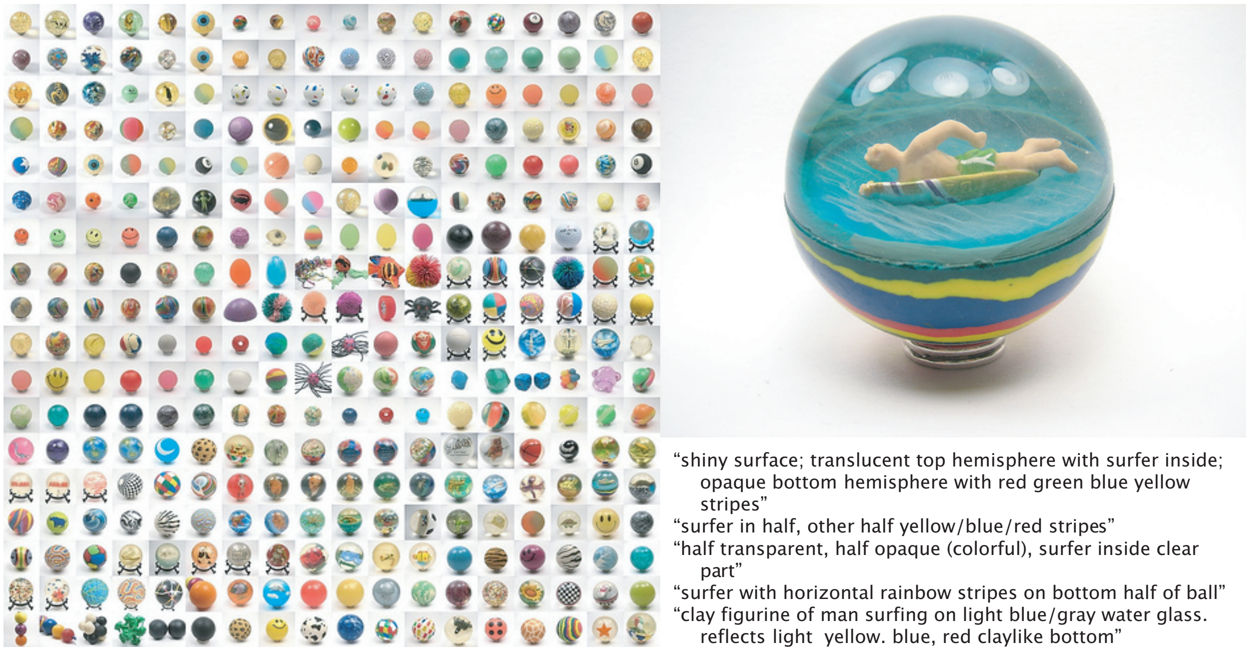
relative to the number of objects with each of the two possible features.

When there were equal numbers of objects with each feature (e.g., two red and two circular objects)—represented by the diagonal in Figure 2—mean bets were very close to \$50, reflecting equal probability. In contrast, in the case shown in Figure 1, there is only one object in the named category (red) but two in the unnamed category (circular). We predicted average bets of \$67, and participants' average response was \$70.

More generally, the correlation between the values predicted by the informative communication model and the experimentally determined means was high (Pearson's $r = .93$, Spearman's rank-order $r = .92$, $p < .0001$). Thus, in their inferences about the meanings of novel words, participants' judgements conformed quite precisely to the predictions of our model.

Experiment 2

In our second experiment, we tested the predictions of the informative communication framework for speakers' productive language use. We used a natural-language description task, rather than probability judgments in a questionnaire. For our stimulus set we chose a set of photos of "superballs"—



"shiny surface; translucent top hemisphere with surfer inside; opaque bottom hemisphere with red green blue yellow stripes"
"surfer in half, other half yellow/blue/red stripes"
"half transparent, half opaque (colorful), surfer inside clear part"
"surfer with horizontal rainbow stripes on bottom half of ball"
"clay figurine of man surfing on light blue/gray water glass. reflects light yellow. blue, red claylike bottom"

Figure 3: Miniature images of the full set of superballs, alongside a ball and descriptions by five participants.

collectible bouncing balls—from an internet photo-sharing website. We selected this stimulus set because each ball had been tagged by the album owner with information about its distinguishing features (providing us with some "ground truth" about word extensions). Crucially, the tags were assigned irrespective of their informativeness relative to this particular set of photos: for example, all 304 photos were assigned the tag "superball." The stimulus set is shown alongside a sample stimulus in Figure 3.

Methods

Participants Forty-four participants from MIT and the surrounding community took part in this experiment as a web survey for which they received payment.

Stimuli and Procedure The stimuli were a set of 304 photos of superballs collected in a Flickr photo album.[3] Each participant was asked to click through the entire set of superballs, one at a time. This process usually took around 5 minutes. Participants next were presented with 50 randomly chosen balls, again one at a time. For each ball in this second set, the participant was instructed to write a short description "so that someone could pick it out of the full set."

Results & Model Fits

Participants' responses were short descriptions of individual superballs. We collected an average of ten descriptions for each ball in the set. Responses varied from verbose descriptions to single-word answers; we treated all words in each description as a "bag of words." Then, to test the hypothesis that the informative communication model provided accurate predictions about participants' productions, we calculated for each ball and tag the probability of a participant writing the corresponding "tag-word" (the word corresponding exactly to the tag) as part of their description of the ball. We then made predictions for the probability of producing a particular tag for each ball using Equation 10. We excluded from our analysis tag-words which were not uttered by any participants in reference to a particular ball and hence may not have been descriptions recognized by participants (e.g., the tag "morris," which referred to the given name of a toy cat pictured in one of the balls).

Across the whole set of tag-words, we found a small but highly significant effect of informativeness on the probability of a tag-word being produced ($r = .19$, $p < .0001$). When we investigated further, we found that the use of basic-level color terms was very poorly predicted by the model ($r = .02$, $p = .63$), and that other terms were much better predicted ($r = .51$, $p < .0001$). Model predictions versus tag-word frequencies words other than color terms are plotted in Figure 4.

Why was the use of particular color words not predicted by our model? There are likely at least three factors that go into the decision to produce a word in a situation like Experiment 2: how informative the word is relative to the other possible words, how frequent or easy to produce the word is (the F term in Equation 6), and how well the word fits the particular object. In order to make a test of the hypothesis that F was the major factor involved in the prediction errors on our model (including its poor performance on colors words), we estimated a measure of F by using the counts of tag-words from the Google Images

[3] http://www.flickr.com/photos/lenore-m/sets/72157600039854497/

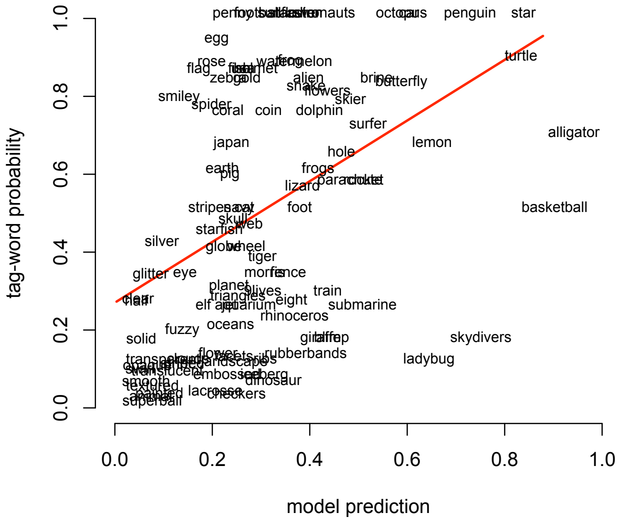


Figure 4: Probability of a label being written in participants' descriptions of a ball, plotted by the predictions of the informative communication model. Points shown are averaged across balls for clearer plotting. Basic level color terms are excluded from this analysis. The line of of best fit is plotted in red.

database (http://images.google.com). Adding this measure to the model predictions only slightly improved the fit to the entire dataset ($r = .22$) and to the non-color words ($r = .52$). We believe that the effects of the third factor—the applicability of tag-words to the images—is likely responsible for the failure of the model to predict the use of color words. As in Figure 3, some color terms applied overall to a particular ball and were used by more participants, while others applied to small parts of the balls and were less widely used.

General Discussion

A model of language as a code for facts does not account for the rich interpretations that language users are able to extract from limited data. Instead, most research on language use in context assumes an intentional inference framework, in which speakers and listeners share assumptions about the nature of the communicative task that allow meanings to be inferred even in the presence of ambiguous or limited data.

Our work here takes a first step towards a formal framework for this kind of inferential account. We used tools from information theory to give a general framework for how speakers can communicate informatively and then used Bayesian inference to derive predictions for both listeners and learners within simple Wittgenstinean "language games." We then tested these predictions in two experiments: a highly constrained word-learning questionnaire and a more natural production experiment. Learners' quantitative judgments were well fit by our model in the first experiment; in the second experiment we found that the model predictions were significantly correlated with speakers' choice of words.

While this framework is related to previous game-theoretic approaches to pragmatics (Benz et al., 2005), it differs from these approaches in that it does not rely on complex, recursive computations but instead on a simple formulation that can be computed whenever the space of meanings and mappings to linguistic expressions is known. With these elements defined, our framework can be used to make predictions in any situation in which a space of possible meanings can be defined, ranging from simple non-linguistic communication experiments to complex cases like scalar implicature and anaphora resolution. Our hope is that future work will make use of this framework to address a broad range of questions in language use and language learning.

Acknowledgments

We thank Pat Shafto and Avril Kenney for valuable discussion. This work supported by a Jacob Javits Graduate Fellowship to the first author and NSF DDRIG #0746251.

References

Benz, A., Jager, G., & Van Rooij, R. (2005). *Game theory and pragmatics*. Palgrave Macmillan.

Clark, H. (1996). *Using language*. Cambridge, UK: Cambridge University Press.

Cover, T., & Thomas, J. (2006). *Elements of information theory*. New York: Wiley-Interscience.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. (in press). Modeling cross-situational word learning through inferences about speakers' referential intentions. *Psychological Science*.

Grice, H. (1975). Logic and conversation. *Syntax and Semantics*, *3*, 41–58.

Piantadosi, S., Goodman, N., Ellis, B., & Tenenbaum, J. (2008). A Bayesian model of the acquisition of compositional semantics. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*.

Shafto, P., & Goodman, N. (2008). Teaching games: Statistical sampling assumptions for learning in pedagogical situations. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*.

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and Cognition*. Oxford, UK: Blackwell Publishers.

Tenenbaum, J., & Griffiths, T. (2002). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–640.

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

Xu, F., & Tenenbaum, J. (2007). Word Learning as Bayesian Inference. *Psychological Review*, *114*, 245.