Testing a rational account of pragmatic reasoning: The case of spatial language

Alexandra Carstensen (abc@berkeley.edu)
Department of Psychology, University of California, Berkeley, CA 94720 USA

Elizabeth Kon (ellie.kon@berkeley.edu)
Cognitive Science Program, University of California, Berkeley, CA 94720 USA

Terry Regier (terry.regier@berkeley.edu)
Department of Linguistics, Cognitive Science Program, University of California, Berkeley, CA 94720 USA

Abstract

How do people recover precise meanings from ambiguous utterances? Frank and Goodman (2012) proposed that listeners do this by rationally combining evidence about word meaning and the salience of particular objects in context. They found that a Bayesian model based on this idea provided a near-perfect account of their empirical data. However, their test of the model was based on communication about simple geometrical objects that varied along only three dimensions. Here, we ask whether their proposal extends to the richer and more complex domain of spatial relations. We find that it does. While the results are not as strong as in their original study, they nonetheless demonstrate that simple formal accounts of communication may capture important aspects of pragmatic inference.

Keywords: Spatial cognition, spatial language, semantics, pragmatics, communication.

Language, thought, and communication

A growing trend in cognitive science views language through the lens of its function: as a vehicle for *informative, efficient communication* (e.g. Piantadosi et al., 2011; 2012; Fedzechkina et al., 2012). One such line of work has argued that systems of word meanings in the world's languages tend to support highly informative communication (Regier et al., 2007; Baddeley & Attewell, 2009; Kemp & Regier, 2012; Khetarpal et al., 2013). However, word meanings necessarily leave much information unspecified—thus, the use of words must be supplemented by pragmatic reasoning to allow speaker and listener to communicate effectively. What principles govern this pragmatic reasoning?

Frank and Goodman (2012) proposed an account of pragmatic reasoning in language use. They argued that in conversation, listeners determine the object to which a speaker is referring by rationally combining two sorts of evidence: one concerning how well the speaker's utterance fits each potential referent, and the other concerning how salient each potential referent is in context. Their study presented evidence that listeners combine these two sources of evidence in accord with Bayes' rule in interpreting the speaker's intention. However, one limitation of their study is that it was based on communication about a very simple and cleanly circumscribed semantic domain, and it is not yet known whether similar results would be obtained in a more complex domain.

We seek to answer that question here, by replicating their study in the context of communication about spatial relations. Languages differ substantially in the ways they partition the spatial domain into semantic categories, and these categories sometimes involve relatively subtle features such as attachment by spiking, or being astraddle, in addition to (from a Western viewpoint) more obvious features such as containment and support (Levinson et al., 2003). Thus, the domain of spatial relations is rich enough to allow a test of Frank and Goodman's (2012) proposal in a semantically complex domain.

We first review Frank and Goodman's study, on which ours is based. We then present our study, which tests their proposal in the spatial domain. To preview our results, we find that their account does predict pragmatic reasoning in the spatial domain, but does not do so as cleanly as in their original study in a simpler domain. We conclude that pragmatic reasoning in more complex domains is substantially but not fully accounted for by their proposal as it stands, and consider possible interpretations of this finding.

Frank & Goodman (2012)

How does a listener interpret a speaker's utterance in context? Imagine that a speaker wishes to refer to a specific referent r_s, which is one of several possible referents in a physical context C, and that the speaker has produced a word w to convey this to a listener. Frank and Goodman (2012) proposed that in such situations, the listener determines the speaker's intended referent through Bayes' rule: [1]

$$(1) \qquad P(r_s \mid w, C) = \frac{P(w \mid r_s, C)P(r_s \mid C)}{\sum_{r' \in C} P(w \mid r', C)P(r' \mid C)}$$

Here, the posterior probability $P(r_s|w,C)$ represents the listener's subjective degree of belief that the speaker's intended referent is r_s, given word w and context C. This quantity is proportional to the product of two terms: (1) the likelihood $P(w|r_s,C)$ of the speaker using word w given that

[1] Our notation differs slightly from that of Frank and Goodman (2012), but there is no difference in intended meaning.

the intended referent was r_s in context C, and (2) the prior probability $P(r_s|C)$ that a word in context C would refer to r_s, without any specification of what that word is. The denominator of Equation 1 is a normalizing constant.

Frank and Goodman (2012) assumed that speakers choose words to be maximally specific—that is, that speakers select the term that picks out the smallest set of possible referents in a given context (cf. Xu & Tenenbaum, 2007). Accordingly, they modeled the likelihood $P(w|r_s,C)$ as:

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

where w is the selected word, $|w|$ is the number of objects in the extension of w, and W is the set of all labels that could be validly applied to the intended referent r_s.

To test their model, Frank and Goodman conducted an experiment with three conditions: one to assess each of the model's three components.[2] In all three conditions, participants viewed a communicative context C consisting of three simple geometrical objects that could vary in shape, color, and pattern. For example, a context might contain a solid blue square, a solid blue circle, and a solid green circle. In the *speaker* condition, one of the objects in the context was highlighted as the intended referent (e.g. the blue circle), and participants were asked to bet on which word (e.g. "blue", "circle") they would use to describe that object in that context; this provides an empirical measure of the likelihood $P(w|r_s,C)$. In the *salience* condition, no object was highlighted—instead, participants were told that a speaker had used an unknown word to refer to one of the objects shown in the context, and they were asked to bet on which object was intended; this provides an empirical measure of the prior $P(r_s|C)$. Finally, in the *listener* condition, participants again saw three objects in context without any object highlighted, but this time were told that a speaker had used a single word (e.g. "blue") to refer to one of the objects, and were asked to bet on which object the speaker intended; this provides an empirical measure of the posterior $P(r_s|w,C)$. Frank and Goodman found that mean bets in the speaker condition were very highly correlated with their model likelihood (Equation 2), and that mean bets in the listener condition were very highly correlated with their model posterior probability (Equation 1). They concluded that this simple model captures "some of the richness of human pragmatic inference in context."

We wished to test whether Frank and Goodman's results generalize to the more complex domain of spatial relations. To that end, we followed their formalization, their experimental design, and their analysis, changing only the character of the stimuli and the words that refer to them, as described below.

[2] We describe their conditions briefly here, and provide concrete examples with spatial stimuli below, when we present our variant of their experiment.

Our study

In our experiment, we replaced Frank and Goodman's simple geometric stimuli with line drawings that depict spatial relations. These were taken from the Topological Relations Picture Series (TRPS; Bowerman & Pederson, 1992), a set of 71 line drawings depicting a variety of spatial relations. Each line drawing shows an orange figure object located relative to a black background object. Figure 1 shows a sample of 10 scenes from the TRPS, categorized according to the spatial naming systems of two languages.

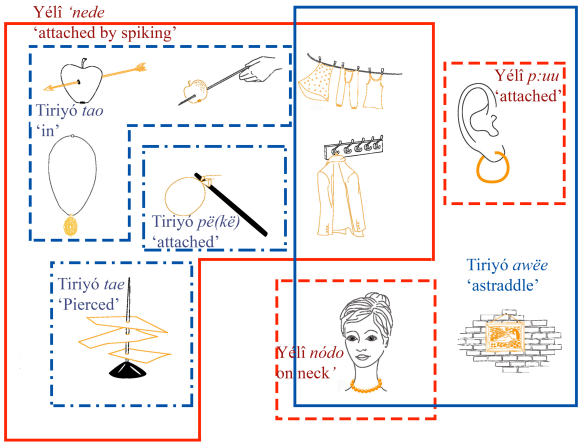

Figure 1: Ten spatial scenes from the TRPS, as categorized in two languages: Tiriyó and Yélî-Dnye. Adapted from Levinson et al. (2003).

The TRPS has been widely used in cross-linguistic studies of spatial language (e.g. Bowerman & Pederson, 1992; Levinson et al., 2003; Khetarpal et al., 2013; Regier et al., 2013), and it represents a broad, rich, and finely-detailed range of different spatial relations. We investigate pragmatic reasoning about reference in the domain of such scenes, using the spatial terms of English (e.g. "in," "around," etc.).

Methods

Participants

A total of 1,427 participants from the U.S. took part in our experiment online through Amazon Mechanical Turk. These participants completed a total of 1,605 trials across all conditions (described below), 447 of which trials were excluded from our analysis because the participant either failed to follow instructions or completed more than one trial, in which case subsequent trials were discarded. Because a unique participant completed each trial in every condition, the number of participants in each condition is equal to the number of trials.

Materials

Communicative contexts were constructed as triads of TRPS scenes presented side by side; an example is shown in Figure 2 below. We divided the 71 scenes of the TRPS into

23 unique triad sets, such that no scene appeared in more than one triad, and we excluded the two remaining scenes (TRPS scenes 2 and 46). Because the goal of the study is to investigate reasoning under ambiguous reference, each triad was formed with the requirement that English spatial terms should be ambiguous when used in the context of the triad. Specifically, every member of the triad shared at least one English spatial term (that could describe that scene) with another member of the triad. Spatial terms for scenes were determined independently of the primary experiment.[3]

Design and procedure

Our design matched that of Frank and Goodman (2012) with one major exception: instead of generating stimuli with pre-determined amounts of referential ambiguity, as they did, we created triads by randomly sampling from our stimulus set with only the above-specified minimal ambiguity requirement within each triad.

There were three conditions, corresponding to the three elements of their model. The speaker condition empirically measured the likelihood $P(w|r_s,C)$; the salience condition empirically measured the prior $P(r_s|C)$; and the listener condition empirically measured the posterior $P(r_s|w,C)$. In all conditions, participants viewed triads of spatial scenes (contexts) and answered questions about them. Figure 2 shows an example trial, with instructions from each of the three conditions.

Speaker (likelihood): Imagine you are talking to someone and you want to refer to the selected scene and distinguish it from the other two scenes. Which word would you use, "**in**" or "**inside**"? Estimate the probability that you would use each word as a percent (responses must add to 100).

Salience (prior) / Listener (posterior): [scene selection highlight not shown] Imagine someone is talking to you and uses [a word you don't know/the word "**inside**"] to refer to one of these scenes. Which scene are they talking about? Estimate the probability that they are talking about each of the scenes as a percent (responses must add to 100).

Figure 2: An example trial, with instructions from the speaker, salience, and listener conditions.

[3] In a separate experiment, 45 UC Berkeley undergraduates, all native English speakers, viewed each TRPS scene and answered the question "Where is the [figure object]?" by completing a fill-in-the-blank sentence that specified figure and ground but not the spatial relationship between them, for instance: "The cup _____ the table." Responses were trimmed to standardize tense and remove non-spatial words (e.g. "is"). To ensure that spatial terms were all of similar complexity, we only included responses with two or fewer spatial morphemes as valid spatial term options. This procedure resulted in each TRPS scene receiving at least two spatial term labels; many received more.

For each of the three conditions, we specify below any elements of procedure not already specified.

Speaker (likelihood). Participants viewed a triad of spatial scenes, one of which was selected as the intended referent. The selected scene (the intended referent) was always indicated by a dotted black square around it. Participants were given a list of all valid spatial terms that could be applied to the intended referent (valid terms were identified independently as described in footnote 3; all other terms were assumed to have 0 probability of applying), and were instructed to estimate the probability that they would use each term in the set to refer to the selected scene, in the context of that triad of scenes. The instructions specified that these probability estimates should add to 100, and this requirement also served as a comprehension check; trials in which participants' estimates did not sum to 100 were discarded and re-run on new participants. Any of the three scenes within a triad could be the selected referent, yielding 3 (scenes per triad) × 23 (triads) = 69 unique trial types (a trial type is a triad with a particular scene selected, as in Figure 2). Scene order was fully counterbalanced within these trial types for a total of 6 orders × 69 trial types = 414 trials in this condition.

Salience (prior). Each participant was shown a triad of spatial scenes without any scene selected, and instructed to imagine that someone had used a word that the participant did not know to refer to one of the scenes in the triad. They were told that this word referred to the scene based on the spatial relationship depicted in it, and were asked to estimate how likely it was that the speaker was referring to each scene, such that their estimates summed to 100. As in the speaker (likelihood) condition, trials in which the participant failed to follow this instruction were discarded and re-run with new participants. Scene order was fully counterbalanced within the triad sets for a total of 6 orders × 23 triads = 138 trials in this condition.

Listener (posterior). Each participant was shown a triad of spatial scenes without any scene selected, together with an English spatial term (e.g. "on") that could be validly applied to at least one scene in the triad. They were asked which scenes in the triad a speaker might be talking about when using that label. Specifically, participants judged how likely it was that each scene was the speaker's intended referent given that spatial term, and entered their judgments as percentages summing to 100. As in the other two conditions, participants whose estimates did not sum to 100 were excluded and the trials re-run. Each of the 23 triads was paired with all possible labels for scenes in that triad, yielding 202 unique trial types. Order within these trial types was pseudo-randomly counterbalanced such that each unique trial type was presented in three of the six possible scene orders, yielding 606 trials in total in this condition.

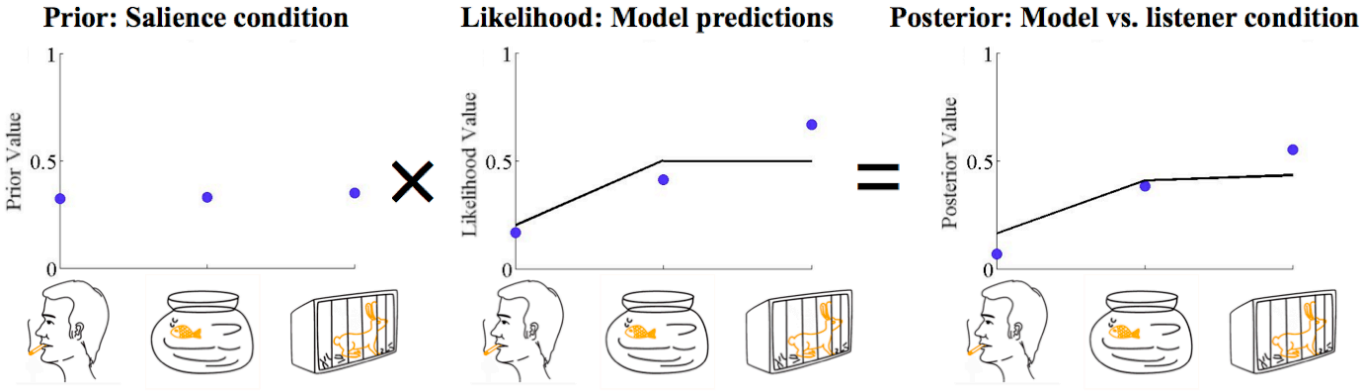

Figure 3: Example model calculation for the triad shown in Figure 2 with the label "inside." Blue dots indicate participants' average responses; black lines indicate model predictions.

Analysis and results

Our analyses followed those of Frank and Goodman (2012). We first tested the model's assumption of speaker informativeness. We then tested whether salience (the prior) predicts responses in the listener condition—to see whether this one source of evidence by itself suffices to explain listeners' inferences. Finally, we assessed the combination of evidence through Bayes' rule, by comparing the model posterior to empirical responses in our listener condition (to which this quantity is intended to correspond), together with a follow-up analysis. Figure 3 illustrates model calculations alongside empirical results for one sample triad of scenes.

Testing the assumption of speaker informativeness

The model likelihood (Equation 2) is based on the assumption that speakers choose words to be maximally informative in context—that is, so that the word chosen will pick out the smallest set of referents possible in a given context. We tested this assumption by comparing empirical data in the speaker (likelihood) condition with the model likelihood term obtained through Equation 2. We found a significant correlation between average empirical likelihoods and model predictions (r = .36, p < .0001; all correlation p-values obtained by permutation test). This result suggests that the model likelihood reasonably approximates speakers' word choice in context, and that speakers do appear to choose their words informatively.

Does salience alone predict listener's inferences?

It is conceivable that listeners might base their judgments of speakers' intentions solely on the salience of particular objects, without reference to how well a given word fits each referent. To test this, we compared empirical data from the salience (prior) condition to empirical data from the listener (posterior) condition. We found no significant correlation (r = .06, p = .09). This means that if the Bayesian model's posterior successfully predicts data from the listener (posterior) condition, that success cannot be due only to the prior, independent of likelihood.

Testing the Bayesian model

Finally, we tested the central claim: that listeners infer speakers' intentions through Bayesian combination of evidence. We combined the empirical prior (from the salience condition) and model likelihood (from Equation 1), to obtain the model's predicted posterior—and compared it to the empirical posterior (listener condition). We found a significant correlation (r = .70, p < .0001; see Figure 4).

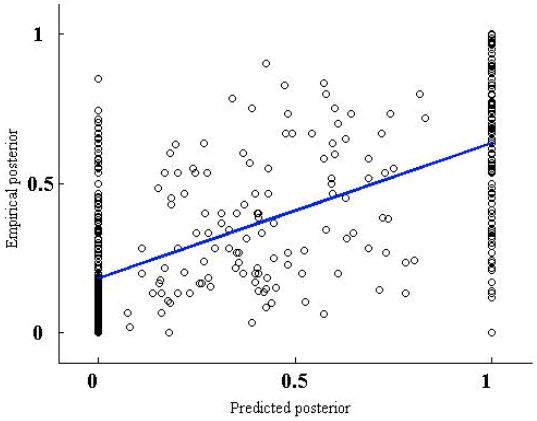

Figure 4: Correlation between model prediction and participants' judgments about which spatial scene the speaker intended, given a speaker's spatial term used in context.

This correlation remains significant when predicted posterior values of one and zero are excluded (r = .36, p < .0001). We also explored another way to obtain a predicted posterior via Bayes' rule: using the empirically determined likelihood (speaker condition) rather than the model likelihood, such that Bayes' rule is now used to combine two empirically determined sources of evidence. In this case, the correlation with the empirical posterior (listener condition) was again significant (r = .70, p < .0001). Thus, Frank and Goodman's (2012) Bayesian account does seem to capture listeners' inference about speakers' intentions

under conditions of referential uncertainty, in a complex and semantically rich domain.

Given these results, and given that we have also found that any success of the Bayesian model cannot be attributed solely to the prior, we sought to understand whether the model's success could be attributed solely to the likelihood instead. To that end, we obtained predictions of listeners' judgments using Equation 1 again, but this time assuming a uniform prior ($P(r_s|C) = ⅓$, $\forall r \in C$), and using the model likelihood of Equation 2. We found that the correlation between this uniform-prior-based model prediction and listener judgments was high ($r = .70$, $p < .0001$)—in fact, it was as high as the correlation we obtained when combining the model likelihood with the empirical prior (salience condition).[4] Thus, it appears that this empirically based prior adds nothing to the predictive power of the model, and the real predictive component is the likelihood. Table 1 summarizes the results of all our analyses, together with analogous analyses by Frank and Goodman (2012).[5]

Table 1: Pearson correlations in the present study compared with those of Frank & Goodman (2012).

Correlation	Present	F&G
Likelihood: model vs. empirical	0.36*	0.98*
Emp. prior vs. emp. posterior	0.06	0.19
Model vs. empirical posterior	0.70*	0.99*
Bayes vs. empirical posterior	0.70*	—
Model with uniform prior vs. empirical posterior	0.70*	—

Discussion and conclusions

We have shown that Frank and Goodman's formalization of pragmatic inference in conversation extends to the domain of spatial relations—a more diverse and naturalistic domain than that of simple geometric objects, in which they originally assessed their proposal. This suggests that their ideas may extend to richer and more complex semantic domains.

However, our results also suggest caution, in at least two respects. First, the correlations between the model prediction and our listener judgment data, while significant, are substantially weaker than those of Frank and Goodman (2012), which were remarkably strong. Frank and Goodman explicitly anticipated that other factors such as word length and frequency—which they provisionally assumed would not be relevant in their initial study—may be relevant more generally; a natural question is whether such factors account for the difference in model fit between our study and theirs. Second, we have seen that the model's success with our data is attributable entirely to the likelihood, and not at all to the prior. It is possible that our empirical prior is, for whatever reason, a flawed measure of the contextual salience of particular objects. This possibility cannot be assumed, but it also cannot be ruled out. Future research can usefully focus on other means of assessing contextual salience, to help resolve this issue.

These caveats notwithstanding, our results do extend Frank and Goodman's (2012) account to a richer and more complex semantic domain, and help to support their conclusion that simple formal accounts of communication may capture important aspects of pragmatic inference.

[4] We also repeated this uniform-prior analysis, but using the empirical (speaker condition) rather than model likelihood, and obtained very similar results ($r = .71$, $p < .0001$).

[5] A possible concern is that the weaker correlations observed in our study are an artifact of the smaller number of observations for each trial (we had 3 or 6 observations whereas Frank & Goodman had 50). However, when we repeated our experiment with 50 observations per trial, we did not see substantial improvements across the reported correlations.

Acknowledgments

We thank Michael Frank, Noah Goodman, Yang Xu, and Joshua Abbott for helpful comments. This work was supported by NSF under grant SBE-1041707, the Spatial Intelligence and Learning Center (SILC), and under NSF Graduate Research Fellowship grant DGE 1106400.

References

Baddeley, R., & Attewell, D. (2009). The relationship between language and the environment: Information theory shows why we have only three lightness terms. *Psychological Science, 20*, 1100-1107.

Bowerman, M., & Pederson, E. (1992). Cross-linguistic studies of spatial semantic organization. In *Annual Report of the Max Planck Institute for Psycholinguistics 1992*, 53-56.

Fedzechkina, M., Jaeger, T., & Newport, E. (2012). Language learners restructure their input to facilitate efficient communication. *PNAS, 109*, 17897–17902.

Frank, M., & Goodman, N. (2012). Predicting pragmatic reasoning in language games. *Science, 336*, 998.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science, 336*, 1049-1054.

Khetarpal, N., Neveu, G., Majid, A., Michael, L., & Regier, T. (2013). Spatial terms across languages support near-optimal communication: Evidence from Peruvian Amazonia and computational analyses. In M. Knauff et al., (Eds.), *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*.

Levinson, S., Meira, S., & the Language and Cognition Group (2003). 'Natural concepts' in the spatial topological domain—Adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. *Language, 79*, 485–516.

Piantadosi, S., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *PNAS, 108*, 3526-3529.

Piantadosi, S., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition, 122*(3), 280-291.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *PNAS, 104*, 1436-1441.

Regier, T., Khetarpal, N., & Majid, A. (2013). Inferring semantic maps. *Linguistic Typology*, *17*, (pp. 89-105).

Xu, F., & Tenenbaum, J. (2007)Word learning as Bayesian inference. *Psychological Review, 114(2)*, 245-272.