On the purpose of ambiguous utterances

Abstract

From the Gricean perspective, ambiguity in language is something to be avoided. However, ambiguity seems ubiquitous during conversations. We propose that speakers sometimes use ambiguous utterances because they are interested in the opinions, preferences, and beliefs of their listeners. We model this hypothesis within the Rational Speech Act framework, adding (i) a mechanism that infers listeners priors when analyzing their responses to (ambiguous) utterances, and (ii) the ability to choose utterances strategically for maximizing information gain. Using an enhanced version of the original reference game experiment (Frank & Goodman, 2012), our model fits participants' data, yielding extremely accurate predictions ($R^2 > 90\%$). This finding suggests that when monitoring a conversational partner's response to an ambiguous utterance, speakers can learn something about the partner's preferences. Moreover, speakers can choose utterances that are expected to yield the most information about those preferences.

Keywords: ambiguity; pragmatics; information gain; predictive priors; Rational Speech Act model

Introduction

Traditionally, linguists have treated ambiguity as a bug in the communication system, something to be avoided or explained away (Grice, 1975; Chomsky, 2002). More recent research has begun to take notice of the efficiency ambiguity affords to us: by relying on context to fill in missing information, we can reuse lightweight bits of language rather than fully specifying the intended message (Levinson, 2000; Piantadosi, Tily, & Gibson, 2012; Wasow, 2015). Viewed in this way, ambiguity serves as a feature—not a bug—of an efficient communication system. This reasoning accords with years of psycholinguistic research documenting that speakers readily produce ambiguous utterances (see Ferreira, 2008, for an overview).

Along related lines, Wasow (2015) reviews a large body of evidence and concludes that ambiguity is rarely avoided, even in situations where it would be communicatively appropriate. This observation stands at odds with the Gricean maxim to avoid ambiguity (Grice, 1975). However, even Grice recognized a case of strategic ambiguity where it could be the intention of the speaker to communicate both possible interpretations afforded by an ambiguous utterance. In such cases, recognition of the ambiguity serves as the communicative purpose of the utterance. Wasow, on the other hand, reviews several cases where ambiguity production serves no obvious communicative purpose.

In search of the communicative purpose of ambiguous language, the current work identifies an additional benefit in using such language: the *extra* information we gain from observing how our listeners resolve ambiguity. We propose that language users learn about each other's private knowledge by observing how they resolve ambiguity. If language does not do the job of specifying the information necessary for full interpretation, then listeners are left to draw on their opinions, beliefs, and preferences to fill in the gaps; by observing how listeners fill those gaps in, speakers learn about the opinions, beliefs, and preferences of the listeners. In a dynamic, naturalistic conversation, speakers can take turns choosing ambiguous statements in order to leave room for their parter to fill the missing information in, thereby revealing opinions, beliefs, and preferences.

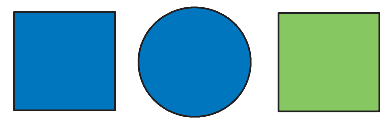

Figure 1: A simple reference game scenario from Frank and Goodman (2012). In the game, speakers choose a single-word utterance to signal one of the objects to a listener. In this scenario, the speaker chooses between the utterances "blue," "green," "square," and "circle."

By way of illustration, take, for example, the scenario in Figure 1. Suppose a speaker produces the single-word utterance "blue" in an attempt to signal one of the objects to a listener. The utterance is ambiguous; the listener can choose either the blue square or the blue circle. Suppose further that, upon hearing "blue," the listener selects the blue circle. In observing this choice, the speaker learns something about the private thoughts of the listener: what made her select the blue circle instead of the blue square? Perhaps the circle is more salient to the listener, or the listener has a preference for circles, or the listener may believe that the speaker has a preference for circles; there may even be a mutual agreement that circles are to be preferred when possible. Importantly, by observing how the listener resolves the ambiguity in reference, the speaker can learn something about the private thoughts of the listener.

However, accessing this added information requires the speaker to reason pragmatically about the pragmatic reasoning of the listener—a higher-order pragmatic reasoning, as it were. In order to select a referent, the listener must interpret the utterance. We follow Frank and Goodman (2012) in treating this interpretation process as active pragmatic, probabilistic reasoning: the listener interprets an utterance by reasoning about the process that generated it, namely the speaker, who selects an utterance by reasoning about how a listener would interpret it. Frank and Goodman model this recursive social reasoning between speakers and listeners, marking the launch of the Rational Speech Act (RSA) modeling framework (see also Franke & Jäger, 2016; Goodman & Frank, 2016).

The current paper builds on the foundational, vanilla RSA model of reference games by introducing uncertainty about the prior beliefs of the listener and modeling a speaker who reasons about these beliefs on the basis of and in anticipation of the observed referent choice. We begin by walking through our modeling assumptions. We then present our model in full detail, and test the behavioral predictions of our model against human data in a series of web-based experiments. We conclude with a discussion of the significance of our findings for understanding ambiguity in natural language, and relate the findings to current theories of predictive coding and active inference.

Model

We begin with the vanilla RSA model of Frank and Goodman (2012). The recursive social reasoning inherent to the RSA modeling framework gets cashed out as various layers of inference. At the base, there is a hypothetical, naïve literal listener L_0 who hears an utterance u and infers the state of the world s that u is meant to describe, that is, the object that may be referred to by the utterance. L_0 performs this inference by conditioning on the literal semantics of u, $[\![u]\!]$. L_0 thus returns a uniform distribution over those states s that can be truthfully described by u:

$$P_{L_0}(s|u) \propto [\![u]\!](s).$$

One layer up, the speaker S_1 observes some state s and chooses an utterance u to communicate that state to L_0. S_1 chooses utterances on the basis of their utility for signaling s to L_0, $U_{S_1}(u;s)$. The speaker's utility maximizes the probability that L_0 would arrive at the correct s on the basis of u, $P_{L_0}(s|u)$, while minimizing the cost of u itself, $C(u)$:

$$U_{S_1}(u;s) = \log(P_{L_0}(s|u)) - C(u).$$

S_1 chooses utterances in proportion to their utility:

$$P_{S_1}(u|s) \propto \exp(\alpha \cdot U_{S_1}(u;s)).$$

At the top layer of inference, the *pragmatic* listener L_1 infers s on the basis of some observed u. The result is a distribution over likely states s; however, unlike L_0, L_1 updates beliefs about the world by reasoning about the process that *generated* u, namely S_1. In other words, L_1 reasons about the s that would have been most likely to lead S_1 to choose the u that was observed:

$$P_{L_1}(s|u) \propto P_{S_1}(u|s) \cdot P(s).$$

Frank and Goodman (2012) tested the predictions of their model against behavioral data from reference games as in Figure 1. To model production behavior (i.e., which utterance should be chosen to communicate a given object) the authors generate predictions from S_1. To model interpretation behavior (i.e., which object the speaker is trying to communicate on the basis of their utterance) the authors generate predictions from L_1. Finding extremely high correlations between model predictions and behavioral data in both cases, Frank and Goodman have strong support for their model of pragmatic reasoning in reference games (see also Qing & Franke, 2015, for a fuller exploration of the modeling choices).

Our model builds on the vanilla version of RSA above by allowing for uncertainty around the listener's state prior, $P(s)$. We have in mind a scenario where a listener might have a preference for a certain object feature (e.g., blue things, squares, circles, etc.), and these preference will influence their object choice. With this in mind, the speaker produces an utterance u, observes the listener's referent choice s, and, on the basis of that choice, infers the preferences f the listener might have had when making the choice. We use the same L_0 and S_1 from the vanilla model. However, we now parameterize L_1's state prior so that it operates with respect to a given feature preference $P(s|f)$:

$$P_{L_1}(s|u,f) \propto P_{S_1}(u|s) \cdot P(s|f).$$

We then model a pragmatic speaker S_2 who updates beliefs about L_1's preferences, $P(f)$. To do so, S_2 produces u, then observes L_1's choice of s, and finally reasons about the likely feature preference f that L_1 used to make that choice:

$$P_{S_2}(f|u,s) \propto P_{L_1}(s|u,f) \cdot P(f).$$

We also model the reasoning process by which a speaker selects the best utterance to learn about the preferences of the listener. Starting with no knowledge of the listener's preferences, S_2 can be assumed to expect a uniform (i.e., flat) feature preference prior $P(f)$. The more the speaker's posterior beliefs about the preferences, $P_{S_2}(f|u,s)$, deviate from the uniform prior, the more the speaker will have learned about the listener's preferences. We can thus model this reasoning in the light of expected information gain, which can be equated with the attempt to maximize the KL divergence between the speaker's flat prior and the expected posterior of the listener's feature preferences f, integrating over all hypothetically possible state observations s:

$$P_b(u) \propto \sum_{s:\ [\![u]\!](s)} \lambda \cdot \mathrm{KL}(P(f), P_{S_2}(f|u,s)) - C(u).$$

We now have two sets of predictions from our model to check: first, the pragmatic speaker's inference about the listener's feature preferences on the basis of their object choice; and second, the pragmatic speaker's strategic utterance selection in the light of the anticipated information gain about the listener's preferences considering their possible object choices. Below, we present two experiments that test these predictions against human behavior.

Experiment 1: Inferring preferences

Our first task is to check the predictions of our S_2 layer: having observed that a listener selects some object s in response to an utterance u, what are the most likely preferences the listener had when making their choice?

Participants

We recruited 90 participants with US IP addresses through Amazon.com's Mechanical Turk crowdsourcing service. Participants were compensated for their participation. On the basis of a post-test demographics questionnaire, we identified

82 participants as native speakers of English; their data were included in the analyses reported below.

Design and methods

We presented participants with a series of reference game scenarios modeled after Figure 1 from Frank and Goodman (2012). Each scenario featured two people and three objects. One of the people served as the speaker, and the other served as the listener. The speaker asks the listener to choose one of the objects, but in doing so she is allowed to mention only one of the features of the target object. Participants were told that the listener might have a preference for certain object features, and participants were tasked with inferring those preferences after observing the speaker's utterance and listener's object choice.

We followed Frank and Goodman (2012) in our stimuli creation. Objects were allowed to vary along three dimensions: color (blue, red, green), shape (cloud, circle, or square), and pattern (solid, striped, polka-dotted). The speaker's utterance was chosen at random from the properties of the three objects present, and the listener's choice was chosen at random from the subset of the three objects that possessed the uttered feature. By varying the object properties, the targeted object, and the utterance, we generated a total of 2400 scenes. Speaker and listener names were chosen randomly in each trial. Participants saw the speaker's utterance in bold (e.g., "green" in Figure 2) and the listener's choice appeared with a dotted orange outline (e.g., the leftmost object in Figure 2). Based on the observed choice, participants were instructed to adjust a series of six sliders to indicate how likely it is that the listener had a preference for a given feature. The sliders specified the six feature values of the two feature dimensions that were not mentioned in the speaker's utterance (e.g., pattern and shape in Figure 2). Participants completed a series of 15 trials. Objects and utterances were chosen as detailed above, with the constraint that 10 trials were potentially informative with respect to listener preferences and 5 trials were uninformative with respect to listener preferences (e.g., observing that the listener chose one of three identical objects).

Results

The S_2 layer of our RSA model returns a posterior distribution over inferred feature preferences f after observing a listener selecting an object in response to an utterance. Model predictions depend on two parameters. The first is the softmax scaling parameter α in the S_1 layer of the model, where the default value is typically set to $\alpha = 1$. This parameter controls how likely S_1 is to maximize utility when choosing utterances. The second parameter scales the strength of individual feature preferences f_i. The softness parameter γ determines how strong a preference is if a certain feature value is indeed preferred. This parameter controls the shape of each of the possible feature preferences f_i that S_2 considers. Preference softness increases with γ: a value of $\gamma = 0$ specifies a hard preference, that is, the listener will always choose the object that contains the preferred feature value if possible. We assume $\gamma = 0$ as the default model value. On the other hand, $\gamma \to \infty$ specifies a uniform feature value preference, that is, no actual preference. As a result, a large value of γ approximates a uniform-prior model.

We optimized α and γ in the light of the KL divergence between the individual participants' slider value choices and the corresponding model predictions:

$$\mathrm{KL} = \sum_{i=1}^{n} P(f_i'|u,s)(\log(P(f_i'|u,s) - \log(P(f_i|u,s)),$$

where $P(f_i'|u,s)$ specifies a participant's normalized slider value settings, that is, empirical estimates of the feature preference posterior given an object scene and a particular utterance u and object choice s; $P(f_i|u,s)$ specifies the respective model prediction value. Since no conclusions can be drawn concerning feature values that are not present in the scene, we ignored the respective feature preference estimates.

By minimizing KL divergence between the empirical and model-predicted preferences for each participant, we maximize the model fit to the individual participants' data. When compared to the uniform-distribution base model, optimizing γ alone yielded a much lower KL divergence value: while the uniform base model yields an average KL divergence of 9.848 (median=7.028), our extended RSA model with individually-optimized γ parameters yielded an average KL divergence of 2.058 (median=1.524). When optimizing α and γ together, a yet smaller average KL divergence of 1.451 (median=0.952) is reached. In the light of the G^2 statistics and under the assumption that we calculated 2 KL divergences in all 15 trials per participant, the KL values should be multiplied by $2*15*2 = 60$ to yield a G^2 estimate, although one may want to consider the two KL divergences in each trial as closely related, such that $2*15 = 30$ may be considered as a more passive multiplier. With this factor, we get a difference of 233.7 between the uniform base model and the one-parameter extended RSA model, while the additional optimization of α improves G^2 further by a value of 18.21. Both of these results are far above the cutoff value of 6.63 for $p = .01$, assuming a Chi-square distribution (Lewandowsky & Farrell, 2011). Thus, the extended model exceeds the uniform base model with very high likelihood. However, considering the much smaller improvement due to the additional optimization of α, we analyze correlations with respect to the one-parameter (i.e., γ-optimized) model in what follows.

When considering the distribution of optimized parameters, we can identify three participants who cause the optimization process to yield γ values above 100, indicating lazy participants who simply leave the slider values unadjusted. Another 15 participants yielded a γ value above 1, which indicates that they are considering preferences only to a small extent. The rest (i.e., 64 participants) yielded values below 1, indicating a strong consideration of preference inferences in line with the model.

To compare RSA model predictions to the human data, we identified all distinguishable cases in the light of the ambigu-

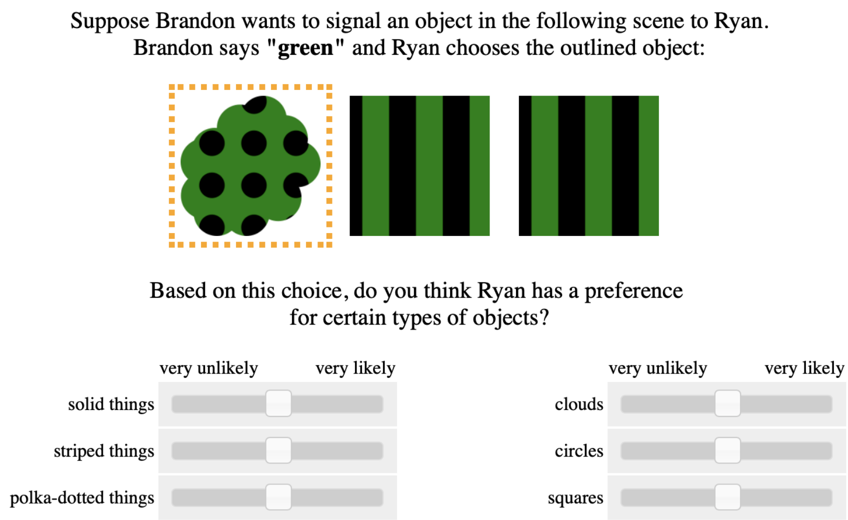

Figure 2: A sample trial from *Experiment 1: Inferring preferences*.

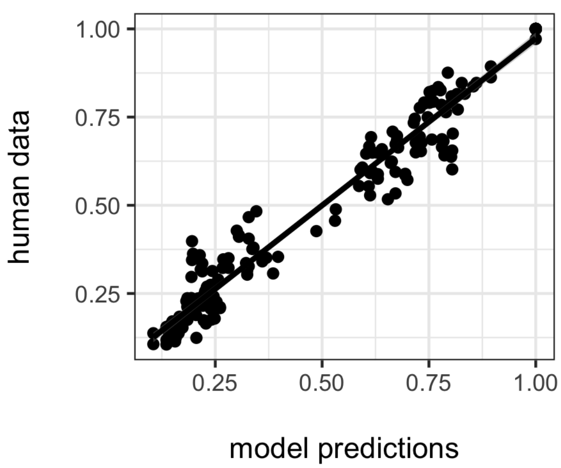

Figure 3: Average human data from Experiment 1 plotted against the predictions of the γ-optimized RSA model; $r^2 = 0.96$, 95% CI [0.94, 0.97].

ity of the chosen utterance and the possible object choices, essentially binning the data of all stimuli for which the extended RSA model yields equal preference-estimate posteriors; there were 48 scene types in our stimuli. For all scene types, the actual feature values and objects in a scene were reordered according to the specific preference inference involved. Thus, after reordering, the results of the individual slider values for individual scenes in each scene type could be averaged for both the participant data and the model predictions.

We plot predictions from the γ-optimized model in Figure 3, where a strong positive correlation between the human judgments and model predictions ($r^2 = 0.96, p < 0.001$) can be observed. Thus, we find strong empirical support for our extended RSA model of preference inference: speakers are indeed able to use listener behavior to arrive at information about their preferences. The question now turns to whether speakers are aware of the relative utility of different utterances when it comes to this reasoning. In other words, are speakers aware that ambiguous language is potentially more informative?

Experiment 2: Choosing utterances

Our next task is to check the predictions of our strategic utterance selection model: given a set of potential referents, which utterance would be most informative with respect to the listener's preferences?

Participants

We recruited 90 participants with US IP addresses through Amazon.com's Mechanical Turk crowdsourcing service; participants in Experiment 1 were not eligible to participate in Experiment 2. Participants were compensated for their participation. On the basis of a post-test demographics questionnaire, we again identified 82 participants as native speakers of English; their data were included in the analyses reported below.

Design and methods

Participants encountered a reference game scenario similar to Experiment 1 in which a speaker signals an object to a listener who might have a preference for certain types of objects. However, rather than observing the utterance and referent choice, participants were now tasked with helping the speaker choose an utterance that was "most likely to reveal the listener's color, shape, or pattern preferences."

We used the same sets of objects from Experiment 1, which could vary along three dimensions. Each trial featured a set of three objects, as in Figure 4. After observing the objects, participants adjusted sliders to indicate which single-feature utterance the speaker should choose. Potential utterances corresponded to the features of the objects present; depending on

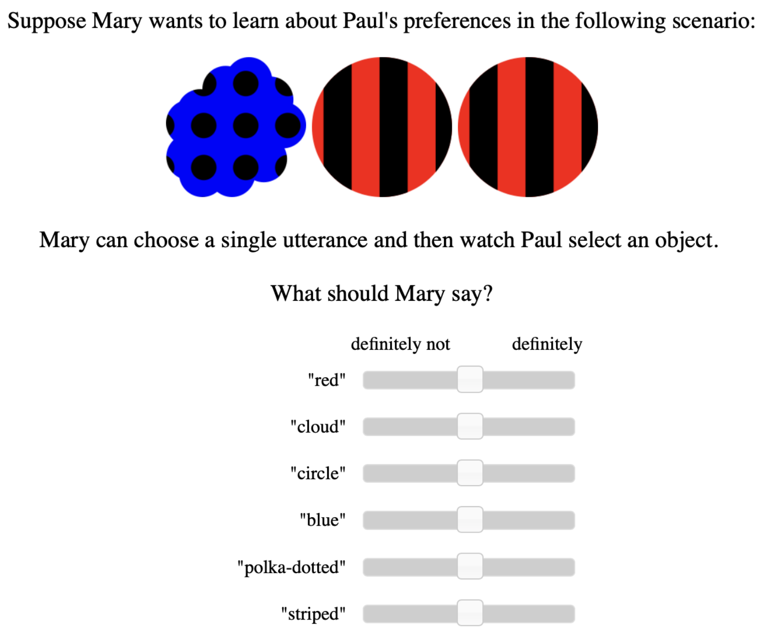

Figure 4: A sample trial from *Experiment 2: Choosing utterances*.

the number of unique features, participants adjusted between three and nine sliders, which specified the present feature values. Participants completed a series of 15 trials. As with Experiment 1, objects were chosen at random, with the constraint that 10 trials were potentially informative with respect to listener preferences (as in Figure 4) and 5 trials were uninformative with respect to listener preferences (e.g., observing a set of three identical objects).

Results

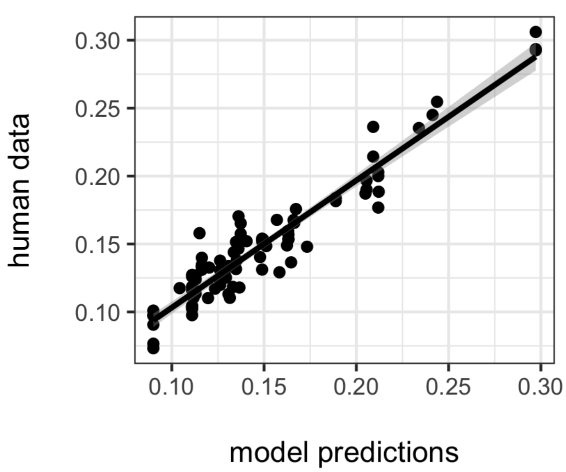

Figure 5: Experiment 2 results; $r^2 = 0.91$, 95% CI [0.84, 0.95].

By reasoning the about predictions of S_2, we are able to use our extended RSA model to compute the expected most informative utterance with respect to inferring preferences. In other words, $P_b(u)$ calculates the probability that a speaker would choose u for the purpose of inferring preferences in our reference game scenario. To generate predictions from $P_b(u)$, a total of three free parameters can be identified. As with the analysis for Experiment 1, we consider different values for α (i.e., speaker's soft-max factor) and γ (i.e., preference softness). We must also set the λ parameter, which factors the importance of choosing the expected most informative utterance with respect to the determined KL divergence values. Note that when allowing negative values for λ, negated information gain essentially minimizes expected information gain. Thus, in this case the model would favor unambiguous utterances. Moreover, when $\lambda = 0$, the model collapses to a uniform distribution of utterance choices, thus yielding a nested model to a uniform base model.

We again optimized model values with respect to KL divergence estimates between the participant data and the model predictions—in this case for utterance preference distributions. The results for the base model yielded an average KL divergence value of 5.415 (median=3.997), while the λ-optimized model yielded a mean KL divergence of 3.774 (median=2.108). Again determining G^2 to enable a Chi-square test for significant model differences in fitting the data, the difference between the base model and one-parameter model (factored by $2 * 15 = 30$) of 49.23 (median difference of 56.67) is highly significant, indicating that the one-parameter model fits the data much better than a uniform base model.

As with the results of the previous experiment, we were able to distinguish three groups of participants on the basis of the fitted parameter values of λ. The first was a "lazy worker" group of 18 participants whose fitted λ values were close to zero (i.e., $-.02 < \lambda < .02$). The second group of 32 participants yielded more negative values (i.e., $-8.20 < \lambda < -.02$), indicating that a significant number of participants preferred to systematically choose non-ambiguous utterances. The third group of again 32 participants yielded more positive values (i.e., $.02 < \lambda < .47$), indicating that these participants indeed chose the most ambiguous utterance in a rather strategic manner.

The proper model fit is confirmed when considering the correlation between the participants' data and the model predictions. As with Experiment 1, we averaged the data and the respective model predictions across specific case bins, which include all scenes that yield identical utterance choice options. In this case, 14 distinct conditions can be identified, with a total of 84 slider values to set. After matching the respective actual feature values with the utterance choice relevant within the respectively-binned conditions, we again computed averages over utterance preference values for the respectively-binned trials across participants and across respective model predictions.

Figure 5 plots predictions from the λ-optimized model, together with the human data. Again, we observe a strong positive correlation between the human judgments and model predictions ($r^2 = 0.91, p < 0.001$). In other words, we find evidence in support of the idea that speakers reason pragmatically about the relative informativity of ambiguous language.

Discussion

We have found strong support for our model of inferring priors on the basis of ambiguous language. The results of Experiment 1 demonstrate that naïve speakers are able to reason pragmatically about *why* listeners may take the actions they do, and the success of our computational model in predicting the observed behavior offers an articulated hypothesis about *how* this reasoning proceeds: when speakers are aware of the ambiguity in their utterances, observing how listeners resolve that ambiguity provides clues to the preferences listeners use when doing so. The results of Experiment 2 demonstrate that speakers are able to capitalize on this reasoning to strategically select utterances that are most likely to inform their understanding of the preferences of their listeners, and that the most informative utterances are also the ambiguous ones.

Taken together, the results of our experiments and the success of our model in predicting those results indicate that humans are aware of the fact that by observing responses to ambiguous utterances, information about the listener's prior preferences can be inferred. Used in this way to inform preferences, ambiguous utterances are closely related to questions, which may ask directly about the relevant preferences. However, ambiguous language provides a ready alternative to asking directly. In normal conversations, a speaker might favor the indirect route afforded by ambiguous utterances, given considerations of politeness and possibly also in an effort to keep the conversation on track.

We note that the analyzed preference prior, viewed from a broader perspective, can be interpreted as the general predictive state of mind of the listener when interpreting the speaker's utterance (Butz & Kutter, 2017). Thus, this state of mind does not only include preferences per se, but all imaginable opinions, beliefs, *and* preferences of the listener. Moreover, during a conversation, this prior will dynamically develop depending on the internal predictive models and the generated utterances and actions of the speaker and listener. This prior depends on the privileged grounds of the conversational partners, and also on the common ground in which the conversation unfolds. Ambiguous utterances are one device for making the privileged ground common knowledge.

References

Butz, M. V., & Kutter, E. F. (2017). *How the mind comes into being: Introducing cognitive science from a functional and computational perspective*. Oxford, UK: Oxford University Press.

Chomsky, N. (2002). An interview on minimalism. In A. Belletti & L. Rizzi (Eds.), *On nature and language* (p. 92-161). Cambridge: Cambridge University Press.

Ferreira, V. S. (2008). Ambiguity, accessibility, and a division of labor for communicative success. *Psychology of Learning and Motivation: Advances in Research and Theory*, *49*, 209-246.

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

Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, *35*(1), 3–44.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818-829.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (p. 26-40). New York: Academic Press.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press.

Lewandowsky, S., & Farrell, S. (2011). *Computational modeling in cognition: Principles and practice*. Thousand Oaks: Sage Publications.

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

Qing, C., & Franke, M. (2015). Variations on a Bayesian theme: Comparing Bayesian models of referential reasoning. In H. Zeevat & H.-C. Schmitz (Eds.), *Bayesian natural language semantics and pragmatics* (p. 201-220). Springer.

Wasow, T. (2015). Ambiguity avoidance is overrated. In S. Winkler (Ed.), *Ambiguity: Laguage and communication* (p. 29-47). de Gruyter.