Social Learning via Ambiguity

December 5, 2019

Abstract

Seemingly in contradiction to Grice's maxim of manner, which postulates that ambiguity should be avoided during conversations to maximize clarity, ambiguity is ubiquitous during conversations. Some ambiguity may be useful for efficiency reasons in cases when clarity is not affected. Here we show that ambiguity can have another, socially highly relevant benefit: responses to ambiguous utterances reveal parts of the internal model of the interpreter. We ran two main types of experiments online using a modified version of the original reference game experiment (Frank & Goodman, 2012) and modeled the recorded response data by enhancing the Rational Speech Act framework. We asked (i) whether speakers can use ambiguity resolving responses as a source of information about the unknown preferences of their conversation partner and (ii) whether speakers are able to strategically chose ambiguous over unambiguous utterances to learn about the preferences of their conversation partner. The experimental and modeling results confirm both points. Participants were able to infer Bayesian posteriors of listeners' preferences when analyzing their choice of objects in a situation of referential ambiguity. Moreover, in the second type of experiment, nearly 40% of the speakers were able to strategically choose ambiguous over unambiguous utterances in an epistemic, goal-directed manner, maximizing expected information gain about the listener's preferences. Surprisingly, an equally large number of participants seemed to minimize expected information gain by systematically choosing unambiguous utterances. Our results thus show that ambiguity resolution can reveal aspects of the knowledge, preferences, and beliefs of conversation partners and some of us are able to strategically use (ambiguous) utterances to gain knowledge about some of these aspects.

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

1 Introduction

The anticipatory nature of the human mind reveals itself in many domains. We anticipate future events and event boundaries and pre-plan our hand movements even before we start moving (?, ?; Belardinelli, Stepper, & Butz, 2016; Belardinelli, Lohmann, Farnè, & Butz, 2018; Lohmann, Belardinelli, & Butz, 2019), revealing anticipatory active inference processes in the sensory-motor domain. In the language domain predictions span across a variety of tasks. Listeners predict the semantic category of upcoming words (Federmeier & Kutas, 2002) as evidenced by a neurophysiological effect known as N400. Comprehension of sentences relies not only on the ability of listeners to anticipate subsequent words based on their transitional probabilities, but also takes into account the structural properties of sentences, revealing an even more abstract level of predictions (Levy, 2008). Dynamic language models show that complex, event-predictive structures guide ambiguity resolution during comprehension and likely also constrain ambiguity generation during language production (?, ?).

These processes have recently been linked to the predictive mind perspective (Friston et al., 2015; ?, ?, ?), proposing that our cognitive mind may be best characterized by an event-predictive inference system (Butz, 2016; Butz & Kutter, 2017). Various discipline associated with cognitive science provide evidence and theoretical considerations in this respect, including neuroscience, linguistics, developmental psychology, and philosophy Zacks:2019,Baldwin:2019,.... The closely related, active inference mechanism (?, ?) implies that our behavior is both, goal-directed and epistemic, expecting increases in internal homeostasis and knowledge gain, respectively. Here, we reveal socially epistemic actions and interaction interpretations, gaining knowledge about the internal models, which are assumed to be used to interpret ambiguous utterances by the conversation partner.

In two main studies, we show how speakers update predictive models of the listener's preferences and beliefs when watching social event interactions, such as when offering a few entities to choose from and observing the choice of the partner. We thus show that humans can interpret behavior of other people as driven by their motives, intentions, or personal characteristics. Conceptually, this idea goes back to the attribution theory (Jones & Davis, 1965; Kelley, 1967; Kelley & Stahelski, 1970). More recently, Shafto, Goodman, and Frank (2012) developed a Bayesian model of learning that formalizes the process of inferring others' knowledge about the world based on their actions and goals.

Here we focus on the situation of ambiguity resolution as a situation that lets the conversation partners observe each other's behavior and reason about internal beliefs that lead to particular choices of interpretations. Moreover, we adapt the Rational Speech Act model framework, reliably modeling the involved, probabilistic interpretation processes as well as epistemic action choice. Interestingly, the modeling work reveals good interpretive abilities but also strong individual differences when the task is to choose (ambiguous) utterances strategically for gaining social knowledge.

The rest of the paper is structured as follows: in Section 2 we first review how different disciplines approached ambiguity in natural language and communication, and then provide a computational background on referential ambiguity resolution. In Section 3, we develop models that are able to infer the preferences of the agent that led her to a particular choice of objects, as well as a model that predicts which utterances are most useful to create the possibility of learning. Sections 4 and 5 demonstrate the results of behavioral experiments, followed by a discussion in Section 6 where we conclude that participants in our experiments were indeed able to use observable behavior of others to infer their prior beliefs, and hypothesize why the ability to intentionally create situations leading to possible learning can be found only in a part of the population.

2 Ambiguity in natural language and communication

2.1 Theoretical approaches

If a speaker and a listener understand an ambiguous utterance differently, communication between them might fail. On rare occasions, such communication failure is deadly. Pinker (2015) alludes to the Charge of the Light Brigade during the Crimean war as an example of a military disaster that was caused by vague orders. He also mentions how poor wording on a warning light was responsible for the nuclear meltdown at Three Mile Island. Finally, citing Cushing (1994), Pinker describes how the deadliest plane crash in history resulted from pilots and air traffic controllers arriving at different interpretations of the phrase 'at takeoff'.

Given that ambiguity can hinder the efficient transfer of information between conversation partners, it is not surprising that linguists have treated the possibility for ambiguity as a bug in the communication system (Grice, 1975; Chomsky, 2002). The attitude towards ambiguity has been quite different in other disciplines though, part of the reason being that the term itself can refer to multiple phenomena. For linguistic research, a word is ambiguous if it can have two separate meanings even in the absence of context, simply as a linguistic sign. In that sense, the word *bat* is ambiguous between a winged mammal and sporting implement. In organizational communication, ambiguity aligns closely with underspecification: an utterance is ambiguous when it does not provide every detail about the intended interpretation, leaving room for the listener to interpret it. In the case of referential ambiguity, an ambiguous utterance may apply to several possible referents in a scene. For example, a pronoun can be referentially ambiguous if there are multiple potential antecedents in the context. It is the latter type that we will be concerned with in this paper.

If we look back at the study of ambiguity, we notice that the strategy of ambiguity avoidance is much older than the pronouncements of modern linguists. Greek and Latin rhetoricians believed that a skillfully written text allows for a perfectly accurate and lossless transmission of meaning to the listener or reader (Ossa-Richardson, 2019); such a text avoids ambiguities.

Still, despite the teachings of classical philologists, authors continued to create ambiguous texts and readers were faced with the challenge of interpreting them. The Bible is one of the most significant of such texts. In the sixteenth century, the Catholic church responded to the Reformation by proposing that the Bible can contain multiple meanings—Ossa-Richardson (2019) equates these meanings with multiple paths that lead readers to God. In a sense, this proposal contained one of the first acknowledgments of the virtue of ambiguity, though with a special caveat—only God could introduce ambiguity, humans should not. The search for efficient transmission of meaning that lasted over millenia rested on an important assumption: we communicate to transfer knowledge to our conversation partner. It is the efficiency of this transfer that many experiments were designed to evaluate. To be more precise, communication was considered efficient if a subject could follow instructions precisely. Yet, ordering actions and following instruction are probably not the most common types of communicative acts (Foppa, 1995) and information-seeking might not be the only communicative task we engage in (Markova & Graumann, 1995)

More recent research has begun to take notice of the efficiency ambiguity affords 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 ambiguous production serves no obvious communicative purpose.

The field of communication sciences views ambiguity as an important communicative tool. In organizational communication—communication that aids production—ambiguity has traditionally stood in opposition to clarity. However, as Eisenberg (1984) notes, clarity is not necessarily a communicative goal in all conversations. Speakers may prefer to remain ambiguous to leave room for the listener's perspective. This freedom is important in communication between managers and their employees, particularly when managers set goals that should stimulate rather than limit future creativity (Mohr, 1983). Ambiguity allows for the expression of ideas that are true of a group of people. For example, consider company slogans or vision statements, where the language must be vague enough to allow every member of the audience to relate to a company's avowed aims (Carmon, 2013). (Pascale & Athos, 1981): interlocutors often employ utterances that allow for a range of interpretations and do not enforce a particular viewpoint.

Eisenberg (1984) further specifies that ambiguity does not necessarily stand in opposition to clarity. In communication with close friends, for instance, interlocutors can use incomplete phrases or vague referential expressions and nevertheless resolve the ambiguity in accordance with the speaker's intention through the use of restricted codes—shared knowledge and beliefs. The participants may not even perceive the utterances as ambiguous in such situations. It is possible that the awareness of a shared code gives rise to the sense of within-group cohesion and social bond between group members. Thus, members of the same group sense a high level of mutual understanding.

2.2 Computational modeling

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.

By way of illustration, take 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 referentially 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 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

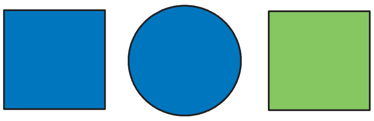


Figure 1: A simple reference game scenario from Frank and Goodman (2012). In the game, speakers choose a single-word utterance to signal one of the objects to a listener. In this scenario, the speaker chooses between the utterances “blue,” “green,” “square,” and “circle.”

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 within the Rational Speech Act (RSA) modeling framework (for more on RSA as a 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 models in full detail, and test the behavioral predictions of our models 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.

3 Model

3.1 Full pragmatic 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 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:

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

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, $P_{L_0}(s \mid u)$[1]:

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

S_1 chooses utterances in proportion to their utility:

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

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:

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

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 presented 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, a pragmatic 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 $P(s \mid f)$:

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

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 and observes L_1's choice of s, then reasons about the likely feature preference f that L_1 used to make that choice:

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

We also model the reasoning process by which a speaker may select 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, $P_{S_2}(f \mid u, s)$, 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

[1] The original model in Frank and Goodman (2012) also includes a term for the cost of utterance $C(u)$. We eliminate the term here since we assume a uniform cost for all the utterances.

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:

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

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.

3.2 Simplified pragmatic model

Sikos, Venhuizen, Drenhaus, and Crocker (2019) demonstrate that simpler models with fewer layers of reasoning perform as well or better than more complex models in simple reference games. In fact, their simplified model with only an enhanced L_0 equipped with an informative prior outperformed the full vanilla RSA model when it came to predicting interpretation behavior, suggesting that participants engage in shallower reasoning than predicted by the full-blown RSA. We therefore compare the predictions of our full pragmatic model with those of a simplified model model.

Our simplified model involves an enhanced literal listener that updates the preference-specific state prior $P(s \mid f)$:

$$P_{L_{0\text{-simp}}}(s \mid u, f) \propto [\![u]\!](s) \cdot P(s \mid f).$$

In the simplified model, the pragmatic speaker $S_{1\text{-simp}}$ reasons directly about the enhanced $L_{0\text{-simp}}$, rather than about L_1 who in turn reasons about S_1's reasoning about L_0.

$$P_{S_{1\text{-simp}}}(f \mid u, s) \propto P_{L_{0\text{-simp}}}(s \mid u, f) \cdot P(f).$$

4 Experiment 1: Inferring preferences

Our first task is to check the predictions of our pragmatic speaker 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?

4.1 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.

4.2 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 possesed 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., "red" in Figure 2) and the listener's choice appeared with a dotted orange outline (e.g., the center 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).

Depending on how many features objects share with the target object (marked by a frame in each trial), we were able to identify 48 ambiguity classes. Ambiguity classes group trials where a model considers a similar number of alternatives that could qualify for the uttered feature. For example, in Figure 2, the utterance *red* picks out 2 possible objects. If, however, the utterance was *green*, only 1 object would qualify, and no learning about preferencs would be possible. In that case, the model would assign equal probability that a person likes dotted objects, striped objects, clouds, or squares. Once the model establishes that more than 1 object can be picked, it also needs to consider whether alternative objects share their features with the target object. For example, if both red objects were also striped, the model would not be able to infer any preferences about the pattern. Finally, we also code whether the objects that were not picked are similar in any of their feature values.

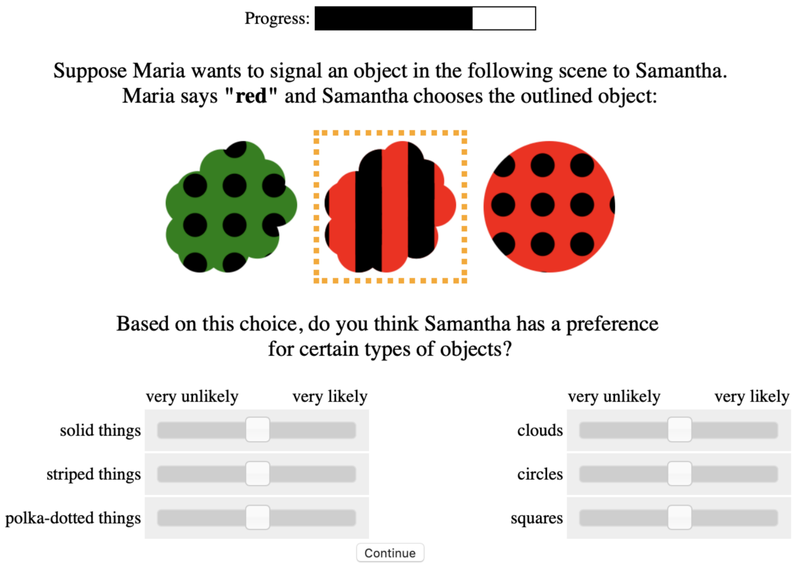
Participants completed a series of 15 trials. Objects and utterances were chosen as detailed above, with the constraint that 10 trials were potentially informative with respect to listener preferences and 5 trials were uninformative with respect to listener preferences (e.g., observing that the listener chose one of three identical objects).

4.3 Results

4.3.1 Models with global optimization

To compare RSA model predictions to the human data, we calculated an average value for each slider binning data into 48 ambiguity classes. We excluded the sliders if their corresponding feature value as not present in a scene. For example, for Figure 2 we excluded the sliders for solid things and squares since none of these are present, and therefore no learning is possible.

We performed two types of optimization: at the individual level and at the global group levels by optimizing a Kullback-Leibler divergence between the data and the model predictions:



Suppose Maria wants to signal an object in the following scene to Samantha.
Maria says "**red**" and Samantha chooses the outlined object:

Based on this choice, do you think Samantha has a preference for certain types of objects?

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

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

where $P(f_i' \mid u,s)$ specifies a participant's normalized slider value settings, which offer empirical estimates of the feature preference posterior given an object scene and a particular utterance u and object choice s; $P(f_i \mid u,s)$ 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. We can then use the KL divergence values to perform the likelihood ratio test for nested models relying on the G^2-statistic which is approximately chi-square distributed (Lewandowsky & Farrell, 2011).

We first present the globally-optimized versions of the model (Figure 4). We fit three parameters for the full RSA model, and two for the simple model. We hypothesized that participants may either go through all the layers of pragmatic reasoning, and additionally calculate the preferences that lead to particular object choice. The last layer of this model S_2 returns a posterior distribution over inferred feature preferences f after observing a listener selecting an object in response to an utterance. In a simpler model, the object choice is driven only by a L_0 semantics enhanced with priors over feature preferences. Upon hearing an utterance *blue* a participants assigns equal probabilities to all blue objects in a scene, and the actual choice of object signals a preference of other feature values that object has. For example, picking a blue circle rather than a blue square is driven by a prefrence for circles.

For the full RSA model, the first parameter is the soft-max scaling factor α in the S_1 layer of the model, where the default value is typically set to $\alpha = 1$ (i.e., no scaling). This parameter controls how likely S_1 is to maximize utility when choosing utterances. The simple model does not rely on this parameter.

The second parameter scales the softness of individual feature preferences f_i. This parameter controls the shape of each of the possible feature preferences f_i that S_2 considers.

$$P(s \mid f) = \frac{P(s \mid f) + \gamma}{\sum P(s \mid f) + 3\gamma}$$

For example, in trial shown in Figure 2, there are two objects that fit the description *red*: a red striped cloud and a red dotted circle. The softness parameter γ regulates the probability that the middle object will be picked if a person has a preference for clouds. Preference softness increases with γ: a value of $\gamma = 0$ specifies a hard preference, which means that the listener will always choose the object that holds the preferred feature value if possible (e.g., clouds when clouds are preferred). We assume $\gamma = 0$ as the default model value. Let us see how the probability changes if γ is set to 0.2:

$$P(s_{\text{cloud}} \mid f_{\text{cloud}}) = \frac{1 + 0.2}{1 + 3 \cdot 0.2} = 0.75$$

Here the preference becomes softer: a subject will pick up clouds with a probability of 0.75 rather than 1. On the other hand, $\gamma \rightarrow \infty$ specifies a uniform feature value preference, or no actual preference. As a result, a large value of γ approximates a uniform-prior model.

Finally, we allow for the possibility of noise in our human data introduced by participants not following instructions. The parameter β, when above 0, allows subjects to consider objects that do not pass the semantic filter of the literal listener.

$$P(s_{\text{x}} \mid u_{\text{x}}) = \frac{P(s \mid u) + \beta}{\sum P(s \mid u) + 3\beta}$$

As β increases, speakers disregard instructions and assign non-zero probabilities to objects that do not correspond to the utterance; with $\beta = 0$, speakers fully obey the instructions and only consider objects with the named property (e.g., only red objects following the utterance *red*).

Figure 3 presents the model performance for the scene type of the sample trial from Figure 2. In that trial, participants saw that the middle object was chosen following the utterance *red*. There are two potential referents for this description: a red striped cloud and a red dotted circle. Since the cloud was chosen, we infer that the person who chose this object has a preference for clouds over circles, and for striped objects vs. dotted ones. From this trial, we cannot learn anything about the preference for solid things or squares, so we expect the subjects to leave the sliders in their default location (i.e., centered). The barplot in Figure 3 shows that indeed when faced with a combination of objects from Figure 2, both humans and the models assign high slider values to clouds and striped things, and low values of the sliders to circles and dotted things. In this plot we show two types of models: an optimized and a non-optimized one.

Both, the simple and the full models with softness optimized globally provide a good linear fit to the data (simple model: $p < 0.001$, $r^2 = 0.86$; full model: $p < 0.001$, $r^2 = 0.86$) suggesting that participants are indeed able to infer the feature preferences that lead to a choice of an object.

4.3.2 Individually-fitted models

Model fit improves when we fit the parameters at the individual level, calculating a parameter estimate for each participant. Individual-level modeling allows us to explore

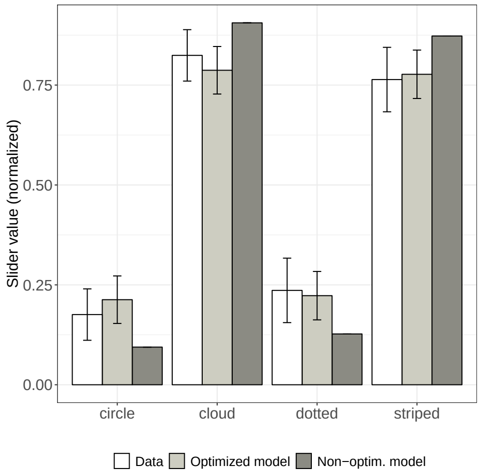


Figure 3: Model predictions and human data for one of the classes of stimuli *Experiment 1: Inferring preferences*.

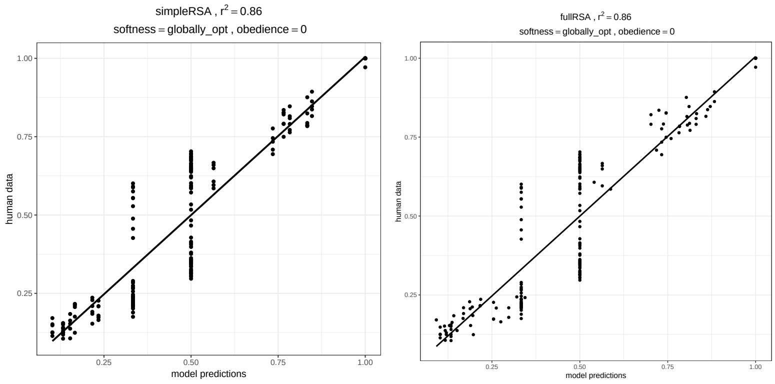


Figure 4: Average human data from Experiment 1 plotted against the predictions of the β-optimized RSA model; both models $r^2 = 0.86$, 95% CI [0.80 0.90].

potential differences between participants, and, more importantly, to evaluate whether the Gricean reasoning strategies apply at the level of individual speakers or only to population as a whole (Franke & Degen, 2016).

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

Figure 5 demonstrates that the full model optimized at the individual level for an additional parameter α does not improve the fit compared to the simplified model. Since the two models account for the same amount of variance in the data, we will proceed with the simplified model evaluation.

We plot predictions from the β and γ-optimized model in Figure 6, where a strong positive correlation between the human judgments and model predictions ($r^2 = 0.99, p < 0.001$) can be observed. The likelihood ratio test revealed that a γ and β-optimized model provides a better fit compared to a model optimized only for γ ($G^2 = 237.36, df = 82, p < 0.01$). The more complex model contains one additional parameter β fitted for

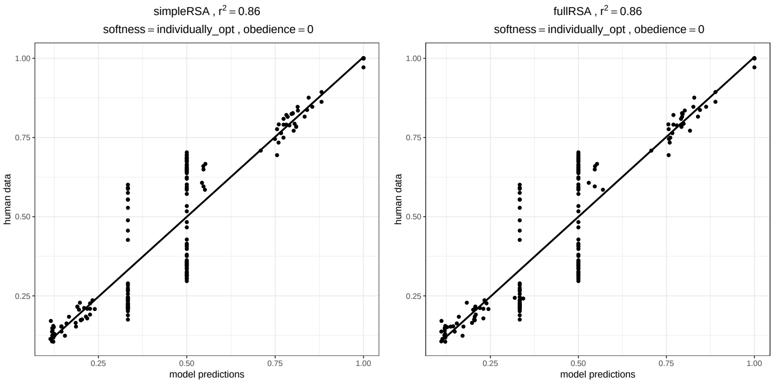


Figure 5: Average human data from Experiment 1 plotted against the predictions of the individually γ-optimized simplified RSA model $r^2 = 0.86$, 95% CI [0.81 0.90] and γ and α optimized full RSA model; $r^2 = 0.86$, 95% CI [0.80 0.90].

each subject giving us 82 degrees of freedom. We additionally check the generalizability of the model by performing a leave-one-out cross-validation. We show in Figure 6 that the cross-validated model retains its fit.

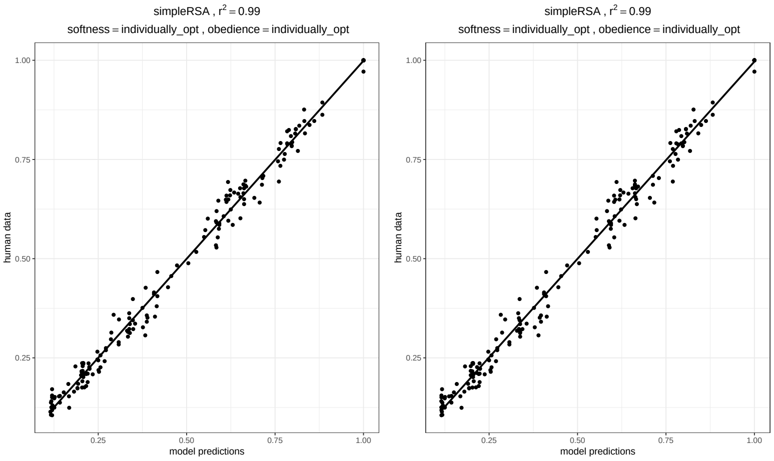


Figure 6: Average human data from Experiment 1 plotted against the predictions of the individually β and γ-optimized simplified RSA model, non-cross-validated (left panel) $r^2 = 0.99$, 95% CI [0.98 1.00] and cross-validated (right panel) $r^2 = 0.99$, 95% CI [0.98 1.01] .

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 able to capitalize on this reasoning when it comes to selecting utterances. In other words, are speakers aware that ambiguous language is potentially more informative?

5 Experiment 2: Choosing utterances

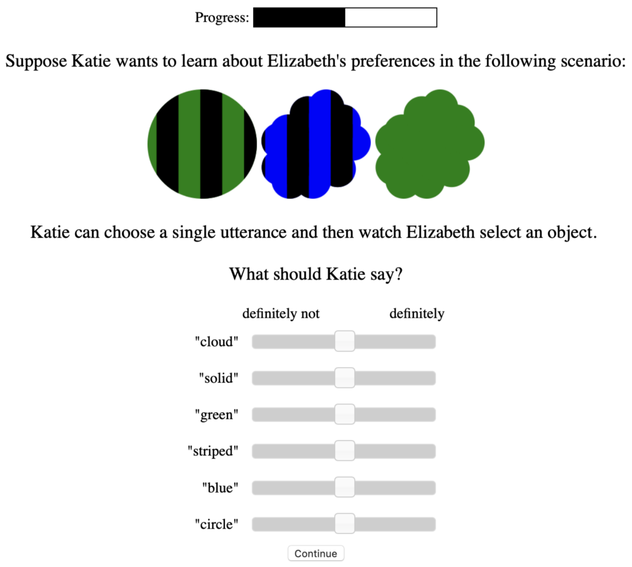
Our next task is to check the predictions of our strategic utterance selection model: given a set of potential referents, are participants able to reason pragmatically about the utility of ambiguous utterances in informing listener preferences?

5.1 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.

5.2 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."



Suppose Katie wants to learn about Elizabeth's preferences in the following scenario:

Katie can choose a single utterance and then watch Elizabeth select an object.

What should Katie say?

definitely not — definitely

"cloud"

"solid"

"green"

"striped"

"blue"

"circle"

Figure 7: A sample trial from *Experiment 2: Choosing utterances*.

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 7. 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 the number of unique features, participants adjusted between three and nine sliders. As with Experiment 1, we averaged the data and the respective model predictions across specific ambiguity classes, 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. Membership within an ambiguity class is defined by how many objects in a scene share each of the features: shape, pattern and color. If objects share a feature, we also consider whether these objects also share other features. For example, in Figure 7, two green objects differ in shape, making the utterance *green* informative. If, on the other hand, both green objects were clouds, uttering *green* would not let the speaker update their belief about listener's preferences. In the most extreme case, when all objects share all three features we are dealing with identical objects. In that situation, all utterances are ambiguous since multiple objects can be picked but no utterance allows the speaker to learn anything about the listener since the object choice is uninformative. Another extreme case is a situation when all objects are unique and do not share any features. Then any utterance will only pick 1 object, making learning about preferences impossible.

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 7) and 5 trials were uninformative with respect to listener preferences (e.g., observing a set of three identical objects).

5.3 Results

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 four 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), and obedience β. 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, with negative values for λ, the model favors unambiguous utterances. Moreover, when $\lambda = 0$, the model collapses to a uniform distribution of utterance choices

Figure 8 shows the model fits of the unfit full and simple RSA models, with both models failing to predict the human data.

To fit model parameters, we again optimized values with respect to KL divergence estimates between the participant data and the model predictions—in this case for utterance preference distributions. We compared three individually-optimized models to determine which model provides the best linear fit to the data. All the models have similar levels of complexity with either softness γ, obedience β, or KL-value factor λ being optimized. The results indicate that we get the best fit by optimizing the KL-factor λ ($r^2 = 0.91$; cross-validated optimization $r^2 = 0.89$), with other models capturing less variance in the data: obedience ($r^2 = 0.80$), softness of preferences ($r^2 = 0.81$). Two- and three-parameter optimization were unstable due to parameter interactions, therefore we do not report the results for those models here.

Unlike for Experiment 1, where even the parameter-free models ensured a good linear fit to the data, optimization produces a large effect on the model predictions in Experiment 2. We illustrate this effect with the predictions of the model for the sample trial (Figure 7). Figure 9 shows that in a situation with a striped green circle, a blue striped cloud, and a solid green cloud, uttering things like *cloud*, *striped*, or *green* and

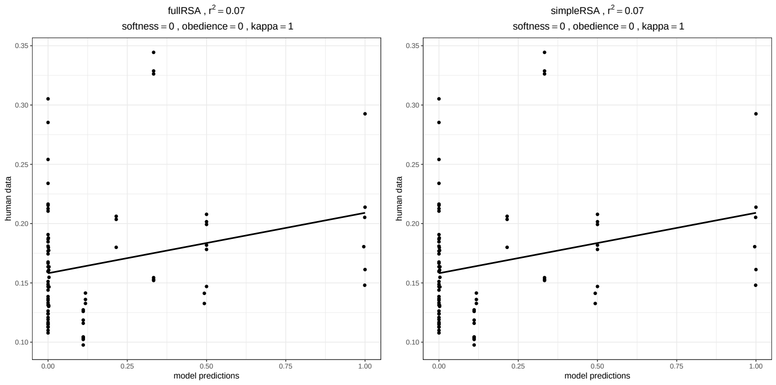


Figure 8: Average human data from Experiment 2 plotted against the predictions of the parameter-free RSA model; both models $r^2 = 0.07$, 95% CI [0.01.23].

then observing the listener pick an referent could let the speaker learn something about the listener's preferences. For example, *green* picks out two objects: a striped green circle and a solid green cloud; after saying "green", the speaker could learn whether the listener prefers striped things over solid things, or circles over clouds.

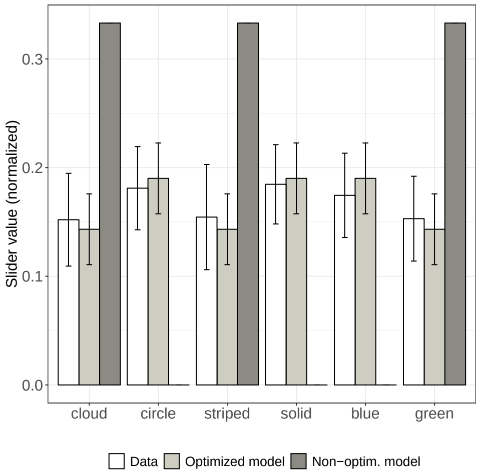


Figure 9: Model predictions and human data for one of the classes of stimuli *Experiment 2: Picking utterances*.

In Figure 10 we compare a λ-optimized model (right panel) to a uniform base model (left panel) that assigns equal probability to each utterance available for a particular context. A model with the λ parameter optimized at the individual level fits the data better than a uniform model (Likelihood ratio test: $G = 268.87, df = 82, p < 0.01$).

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., $-7.13 < \lambda < -.02$), indicating that a significant number of participants preferred to systematically choose unambiguous

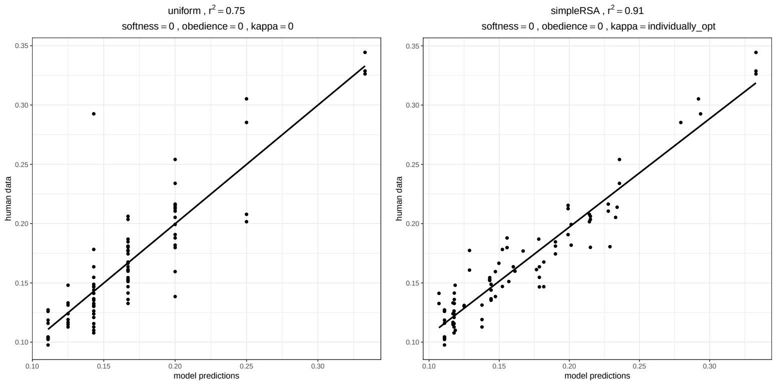


Figure 10: Average human data from Experiment 2 plotted against the predictions of the uniform and simple RSA models; uniform model $r^2 = 0.75$, 95% CI [0.65 0.84], KL-factor optimized $r^2 = 0.91$, 95% CI [0.92 1.06]].

utterances. The third group of again 32 participants yielded more positive values (i.e., $.02 < \lambda < .54$), indicating that these participants indeed chose the most ambiguous utterance in a strategic manner.

6 General 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 at least some speakers are able to capitalize on this reasoning to strategically select ambiguous utterances that are most likely to inform their understanding of the preferences of their listeners.

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 open, in that the conversation partner can choose to disambiguate the ambiguous utterance or, alternatively, choose to continue in a different direction or even to change topic.

We note that the analyzed preference prior, viewed from a broader perspective, can be closely related to a part of the predictive mind of the listener and the speaker (Butz, 2016; Butz & Kutter, 2017). When interpreting an utterance – in our case opening up a set of choices – the listener's mind infers the current choices and integrates them with the preference priors, implicitly anticipating possible choice consequences. Moreover, the expected information gain term – computing the utterance choice of the speaker – can be equated with the computation of socially-motivated active inference (Butz, 2017; Friston et al., 2015). It causes the model to strive for an anticipated epistemic value that quantifies the expected information gain about the assumed preference priors of the listener, that is, expected social information gain.

In general, predictive states of mind about others do not only include considerations of the preferences of others, but generally may concern all imaginable knowledge, opinions, beliefs, current train of thought considerations, and preferences of the listener. Moreover, during a conversation, the involved "social" priors will dynamically develop depending on the internal predictive models and the generated utterances, actions, and responses of the speaker and listener. The prior dynamically depend on the privileged grounds of the conversational partners, and also on the common ground in which the conversation unfolds. In that sense, ambiguous utterances are one device for projecting parts of each others' privileged grounds into the common ground.

References

Belardinelli, A., Lohmann, J., Farnè, A., & Butz, M. V. (2018). Mental space maps into the future. *Cognition*, *176*, 65–73.

Belardinelli, A., Stepper, M. Y., & Butz, M. V. (2016). It's in the eyes: Planning precise manual actions before execution. *Journal of vision*, *16*(1), 18–18.

Butz, M. V. (2016). Towards a unified sub-symbolic computational theory of cognition. *Frontiers in Psychology*, *7*(925). doi: 10.3389/fpsyg.2016.00925

Butz, M. V. (2017). Which structures are out there? learning predictive compositional concepts based on social sensorimotor explorations. In T. K. Metzinger & W. Wiese (Eds.), . Frankfurt am Main: MIND Group. Retrieved from http://predictive-mind.net/papers/which-structures-are-out-there doi: 10.15502/9783958573093

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.

Carmon, A. F. (2013). Is it necessary to be clear? an examination of strategic ambiguity in family business mission statements. *Qualitative Research Reports in Communication*, *14*(1), 87–96. Retrieved from https://doi.org/10.1080/17459435.2013.835346 doi: 10.1080/17459435.2013.835346

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

Cushing, S. (1994). *Fatal words: Communication clashes and aircraft crashes*. University of Chicago Press.

Eisenberg, E. M. (1984). Ambiguity as strategy in organizational communication. *Communication monographs*, *51*(3), 227–242.

Federmeier, K. D., & Kutas, M. (2002). Picture the difference: Electrophysiological investigations of picture processing in the two cerebral hemispheres. *Neuropsychologia*, *40*(7), 730–747.

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.

Foppa, K. (1995). On mutual understanding and agreement in dialogues. In *Mutualities in dialogue.* Cambridge, UK: Cambridge University Press.

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

Franke, M., & Degen, J. (2016). Reasoning in reference games: Individual-vs. population-level probabilistic modeling. *PloS one*, *11*(5), e0154854.

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

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., FitzGerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, *6*, 187-214. doi: 10.1080/17588928.2015.1020053

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.

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions the attribution process in person perception. In *Advances in experimental social psychology* (Vol. 2, pp. 219–266). Elsevier.

Kelley, H. H. (1967). Attribution theory in social psychology. In *Nebraska symposium on motivation.*

Kelley, H. H., & Stahelski, A. J. (1970). Social interaction basis of cooperators' and competitors' beliefs about others. *Journal of personality and social psychology*, *16*(1), 66.

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

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

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

Lohmann, J., Belardinelli, A., & Butz, M. V. (2019). Hands ahead in mind and motion: Active inference in peripersonal hand space. *Vision*, *3*(2), 15.

Markova, I., & Graumann, F. K., Carl F. (1995). Preface. In I. Markova & F. K. Graumann Carl F. (Eds.), *Mutualities in dialogue.* Cambridge, UK: Cambridge University Press.

Mohr, L. B. (1983). The implications of effectiveness theory for managerial practice in the public sector. In K. S. Cameron & D. A. Whetten (Eds.), *Organizational effectiveness* (pp. 225–239). Elsevier.

Ossa-Richardson, A. (2019). *A history of ambiguity*. Princeton University Press.

Pascale, R. T., & Athos, A. G. (1981). *The art of Japanese management*. New York: Simon & Schuster.

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

Pinker, S. (2015). *The sense of style: The thinking person's guide to writing in the 21st century*. Penguin Books.

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.

Shafto, P., Goodman, N. D., & Frank, M. C. (2012). Learning from others: The consequences of psychological reasoning for human learning. *Perspectives on Psychological Science*, *7*(4), 341–351.

Sikos, L., Venhuizen, N., Drenhaus, H., & Crocker, M. (2019, 04). *Reevaluating pragmatic reasoning in web-based language games.* doi: 10.13140/RG.2.2.30535.14249

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