Optimal Reasoning About Referential Expressions

Judith Degen
Dept. of Brain and Cognitive Sciences
University of Rochester
jdegen@bcs.rochester.edu

Michael Franke
ILLC
Universiteit van Amsterdam
m.franke@uva.nl

Abstract

The *iterated best response (*IBR*) model* is a game-theoretic approach to formal pragmatics that spells out pragmatic reasoning as back-and-forth reasoning about interlocutors' rational choices and beliefs (Franke, 2011; Jäger, 2011). We investigate the comprehension and production of *referential expressions* within this framework. Two studies manipulating the complexity of inferences involved in comprehension (Exp. 1) and production (Exp. 2) of referential expressions show an intriguing asymmetry: comprehension performance is better than production in corresponding complex inference tasks, but worse on simpler ones. This is not predicted by standard formulations of IBR, which makes categorical predictions about rational choices. We suggest that taking into account quantitative information about beliefs of reasoners results in a better fit to the data, thus calling for a revision of the game-theoretic model.

1 Introduction

Reference to objects is pivotal in communication and a central concern of linguistic pragmatics. If interlocutors were ideal reasoners, speakers would choose the most convenient referential expression that is sufficiently discriminating given the hearer's perspective, while hearers would choose the referent for which an observed referential expression is optimal given the speaker's perspective. But it would be folly to assume that humans are ideal reasoners, so the question is: how much do interlocutors take each other's perspective into account when producing and interpreting referential expressions?

A lot of work has been dedicated to this issue. For example, computational linguists have investigated efficient and natural rules for generating and comprehending referential expressions (see Dale and Reiter (1995) and Golland et al. (2010) for work directly related to ours). Many empirical studies have addressed the more specific questions of whether, when and/or how, hearers take speakers' *privileged information* into account (Keysar et al., 2000; Keysar et al., 2003; Hanna et al., 2003; Heller et al., 2008; Brown-Schmidt et al., 2008). Also, eye-tracking studies in the visual-world paradigm have been used to investigate how *quantity reasoning* influences the interpretation of referential expressions (Sedivy, 2003; Grodner and Sedivy, 2011; Huang and Snedeker, 2009; Grodner et al., 2010). In recent work closely related to ours, Stiller et al. (2011) and Frank and Goodman (2012) proposed a Bayesian model of producing and comprehending referential expressions in a game setting similar to the kind we consider here. We will more closely compare these related approaches in Section 6. Despite these various efforts, it is still a matter of debate whether or to what extent interlocutors routinely consider each other's perspective.

In order to contribute to this question, we follow a recent line of experimental approaches to formal epistemology and game theory (Hedden and Zhang, 2002; Crawford and Iriberri, 2007) to investigate how much *strategic* back-and-forth reasoning speakers and hearers employ in abstract language games. The tasks we investigate translate directly to the kind of *signaling games* that have variously been used to account for a number of pragmatic phenomena, most notably *conversational implicatures* (see, e.g., Parikh (2001), Benz and van Rooij (2007) or Jäger (2008)). A benchmark model of idealized step-by-step reasoning, called *iterated best response (*IBR*) model*, exists for these games (Franke, 2011; Jäger, 2011). IBR makes concrete predictions about the depth of strategic reasoning required to "solve" different kinds of referential language games, so that by varying the difficulty of our referential tasks, it is possible to both: (i) test the predictions of IBR models of pragmatic reasoning and (ii) determine the extent to which speakers and hearers reason strategically about the use of referential expressions.

Our data shows that participants perform better at reasoning tasks that IBR predicts to involve fewer inference steps. This holds for comprehension and production. However, our data also shows an interesting asymmetry: comprehension performance is better than production in corresponding complex inference tasks, but worse on simpler ones. This is not predicted by standard formulations of IBR which makes categorical predictions about rational choices. However, it is predicted by a more nuanced variation of IBR that pays attention to the quantitative information in the belief hierarchies postulated by the model.

Section 2 introduces signaling games as abstract models of referential language use. Section 3 outlines the relevant notions of IBR reasoning. Sections 4 & 5 describe our comprehension and production studies respectively. Section 6 discusses the results.

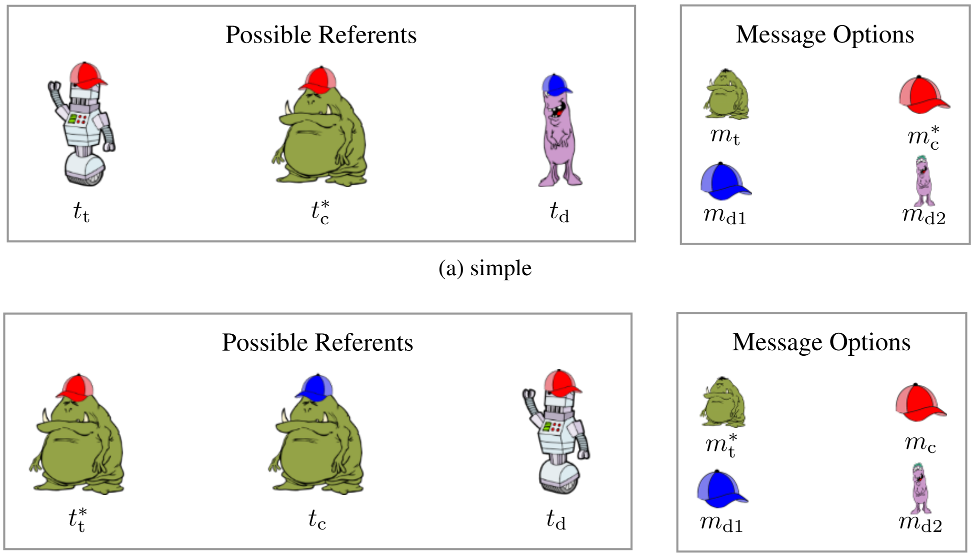
2 Referential Language Games

If speaker and hearer share a commonly observable set T of possible referents in their immediate environment, referential communication has essentially the structure of a *signaling game*: the sender S knows which $t \in T$ she wants to talk about, but the receiver R does not; the speaker chooses some description m; if R can identify the intended referent, communication is successful, otherwise a failure. Such a game consists of a set T (of possible referents), a set M of messages that S could use, a prior probability distribution $\Pr$ over T that captures R's prior expectation about the most likely intended referent, and a utility function that captures the players' preferences in the game. We assume that S and R are both interested in establishing reference, so that if t is the intended referent and t' is R's guess, then for some constants $s > f$: $\mathrm{U}(t, t') = s$ if $t = t'$ and f otherwise. Additionally, if messages are meaningful, this is expressed by a denotation function $[\![m]\!] \subseteq T$ that gives the set of referents to which m is applicable (e.g., of which it is true).

Consider, e.g., the situations depicted in Fig. 1. There are three possible referents $T = \{t_\mathrm{t}, t_\mathrm{c}, t_\mathrm{d}\}$ in the form of monsters and robots wearing one accessory each that both S and R observe. Since there is no reason to prefer any referent over another, we assume that $\Pr$ is a *flat* distribution over T. There are also four possible messages $M = \{m_\mathrm{t}, m_\mathrm{c}, m_\mathrm{d1}, m_\mathrm{d2}\}$ with some intuitively obvious "semantic meaning". For example, the message m_c for *red hat* would intuitively be applicable to either the *robot* t_t or the *green monster* t_c, so that $[\![m_\mathrm{c}]\!] = \{t_\mathrm{t}, t_\mathrm{c}\}$.

Signaling games like those in Fig. 1 are the basis for the critical conditions of our experiments (see also Sections 4 and 5), where we test which referent subjects choose for a given *trigger message* and which message they choose for a *trigger referent*. Trigger items for comprehension and production experiments are marked with an asterisk in Fig. 1. Indices t, c, d stand for *target*, *competitor* and *distractor* respectively.

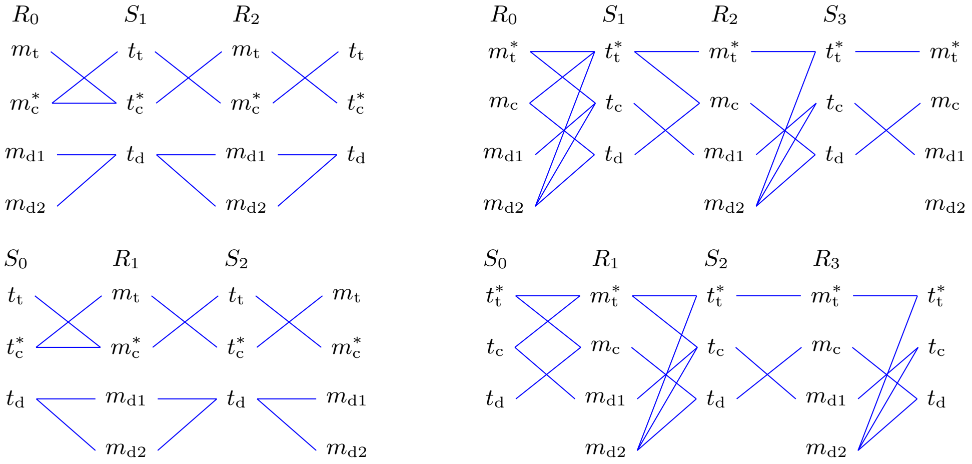
We refer to a game as in Fig. 1(a) as the *simple implicature condition*, because it involves a simple *scalar implicature*. Hearing *trigger message* m^*_c, R should reason that S must have meant *target state* t_t, and not *competitor state* t_c, because if S had wanted to refer to the latter she could have used an unambiguous message. Conversely, when S wants to refer to *trigger state* t^*_c, she should not use the true but semantically ambiguous message m_c, because she has a stronger message m_t. Similarly, we refer to a game in Fig. 1(b) as the *complex implicature condition*, because it requires performing scalar reasoning twice in sequence (see Fig. 2 later on).



(a) simple

(b) complex

Figure 1: Target implicature conditions. Hearers choose one of the POSSIBLE REFERENTS $T = \{t_\mathrm{t}, t_\mathrm{c}, t_\mathrm{d}\}$. Speakers have MESSAGE OPTIONS $M = \{m_\mathrm{t}, m_\mathrm{c}, m_\mathrm{d1}, m_\mathrm{d2}\}$. Trigger items are indicated with asterisks: e.g., t_t^* is the referent to be communicated on complex production trials.



(a) simple

(b) complex

Figure 2: Qualitative predictions of the IBR model for simple and complex conditions. The graphs give the set of best responses at each level of strategic reasoning as a mapping from the left to the right.

3 IBR Reasoning

The IBR model defines two independent strands of strategic reasoning about language use: one that starts with a naïve (level-0) receiver R_0 and one that starts with a naïve sender S_0 (Franke, 2011; Jäger, 2011). If utilities are as indicated and priors are flat, the behavior of level-0 players is predicted to be a uniform choice over options that conform to the semantic meaning of messages: $R_0(m) = [\![m]\!]$ and $S_0(t) = \{m \mid t \in [\![m]\!]\}$. Sophisticated player types of level $k + 1$ play any rational choice with equal probability given a belief that the opponent player is of level k. For our experimental examples, the "light" system of Franke (2011) applies, where sophisticated types are defined as:[1]

$$S_{k+1}(t) \;= \begin{cases} \arg\min_{m \in R_k^{-1}(t)} \mid R_k(m) \mid & \\ \qquad\qquad \text{if } R_k^{-1}(t) \neq \emptyset & \\ S_0(t) \qquad \text{otherwise} & \end{cases}$$

$$R_{k+1}(m) = \begin{cases} \arg\min_{t \in S_k^{-1}(m)} \mid S_k(t) \mid & \\ \qquad\qquad \text{if } S_k^{-1}(m) \neq \emptyset & \\ R_0(m) \qquad \text{otherwise} & \end{cases}$$

The sequences of best responses for the simple and complex games from Fig. 1 are given in Fig. 2. On this purely qualitative picture, the IBR model makes the same predictions for comprehension and production. In the simple condition, the trigger item is mapped to either target or competitor with equal chance by naïve players; all higher level types map the trigger item to the target item with probability one. In the complex condition, the trigger items are mapped to target and competitor in levels 0 and 1 with equal probability, but uniquely to the target item for $k \geq 2$.

The sequences in Fig. 2 only consider the actual best responses of S and R, but not the more nuanced quantitative information that gives rise to these. Best responses are defined as those that maximize expected utility given what the players believe about how likely each choice option would lead to communicative success. The relevant expected success probabilities are given in Table 1 for sophisticated types. (Naïve types have no or only trivial beliefs about the game.)

For reasons of space suffice it to give the intuition behind these numbers. E.g., in the simple condition R_1 believes that the trigger message is used by naïve senders who want to refer to t_t or t_c. But naïve senders who want to refer to t_c would also use m_t with probability 1/2. So, by Bayesian conditionalization, after hearing m_c, R_1 believes the intended referent is t_t with probability 2/3.

Notice that while R's success expectations always sum to one (there is always only exactly one intended referent), S's success expectations need not (several messages could be believed to lead to successful communication). A further difference concerns when S and R are sure of communicative success. In the simple condition, S_1 is already sure of success, but only $R_{\geq 2}$ is. In the complex condition, R_2 is already sure of success, but only $S_{\geq 3}$ is. So, if we assume that human reasoners aim for certainty of communicative success in pragmatic reasoning, the simple condition is less demanding in production than in comprehension, while for the complex condition the reverse is the case.

4 Experiment 1

Exp. 1 tested participants' behavior in a *comprehension* task that used instantiations of the signaling games described in Section 2.

4.1 Methods

Participants. Using Amazon's Mechanical Turk, 30 workers were paid \$0.60 to participate. All were naïve as to the purpose of the experiment and participants' IP address was limited to US addresses only. Two participants did the experiment twice. Their second run was excluded.

Procedure and Materials. Participants engaged in a referential comprehension task. On each trial they saw three objects on a display. Each object differed systematically along two dimensions: its ontological kind (robot or one of two monster species) and accessory (scarf or either blue or red hat). In addition to these three objects, participants saw a pictorial message that they were told was sent to them by a previous participant whose job it was to get them to pick out one of these three objects. They

[1] Here $R_k^{-1}(t) = \{m \mid t \in R_k(m)\}$. Likewise for S_k^{-1}.

	simple		complex	
level	R	S	R	S
1	$\langle 2/3, 1/3, 0\rangle$	$\langle 1, 1/2, 0, 0\rangle$	$\langle 1/2, 1/2, 0\rangle$	$\langle 1/2, 1/2, 0, 1/3\rangle$
2	$\langle 1, 0, 0\rangle$	$\langle 1, 0, 0, 0\rangle$	$\langle 1, 0, 0\rangle$	$\langle 1/2, 0, 0, 1/3\rangle$
3	$\langle 1, 0, 0\rangle$	$\langle 1, 0, 0, 0\rangle$	$\langle 1, 0, 0\rangle$	$\langle 1, 0, 0, 1/3\rangle$

Table 1: Success expectations for the trigger items in the simple and complex condition. Success expectations for R are given in order for t_{t}, t_{c} and t_{d}, those for S in order for m_{t}, m_{c}, m_{d1} and m_{d2}.

were told that the previous participant was allowed to send a message expressing only one feature of a given object, and that the messages the participant could send were furthermore restricted to monsters and hats. The four expressible features were visible to participants at the bottom of the display on every trial.

Participants initially played four sender trials. They saw three objects, one of which was highlighted with a yellow rectangle, and were asked to click on one of four pictorial messages to send to another Mechanical Turk worker to get them to pick out the highlighted object. They were told that the other worker did not know which object was highlighted but knew which messages could be sent. The four sender trials contained three unambiguous and one ambiguous trial which functioned as fillers in the main experiment.

Participants saw 36 experimental trials, with a 2:1 ratio of fillers to critical trials. Of the 12 critical trials, 6 constituted a simple implicature situation and 6 a complex one as defined in Section 2 (see also Fig. 1).

Target position was counterbalanced (each critical trial occurred equally often in each of the 6 possible orders of target, competitor, and distractor), as were the target's features and the number of times each message was sent. Of the 24 filler trials, half used the displays from the implicature conditions but the target was either t_{c} or t_{d} (as identified unambiguously by the trigger message). This was also intended to prevent learning associations of display type with the target. On the other 12 filler trials, the target was either entirely unambiguous or entirely ambiguous given the message. That is, there was either only one object with the feature denoted by the trigger message, or there were two identical objects that were equally viable target candidates. Trial order was pseudo-randomized such that there were two lists (reverse order) of three blocks, where critical trials and fillers were distributed evenly over blocks. Each list began with three filler trials.

4.2 Results and Discussion

Proportions of choice types are displayed in Fig. 3(a). As expected, participants were close to ceiling in choosing the target on unambiguous filler trials but at chance on ambiguous ones. This confirms that participants understood the task. On critical implicature trials, participants' performance was intermediate between ambiguous and unambiguous filler trials. On simple implicature trials, participants chose the target 79% of the time and the competitor 21% of the time. On complex implicature trials, the target was chosen less often (54% of the time).

To test whether the observed differences in target choices above were significantly different, we fitted a logistic mixed-effects regression to the data. Trials on which the distractor was selected were excluded to allow for a binary outcome variable (target vs. no target choice). This led to an exclusion of 5% of the data. The model predicted the log odds of choosing a target over a competitor from a Helmert-coded CONDITION predictor, a predictor coding the TRIAL number to account for learning effects, and their interaction. Three Helmert contrasts over the four relevant critical and filler conditions were included in the model, comparing each condition with a relatively less skewed distribution against the more skewed distributions (in order: ambiguous fillers, complex implicatures, simple implicatures, unambiguous fillers). This allowed us to capture whether the differences in distributions for neighboring conditions suggested by Fig. 3(a) were significant. We included the maximal random effect structure that allowed the model to converge:[2] by-participant ran-

[2] For the procedure that was used to generate the random effect structure, see http://hlplab.wordpress.com/

	Coef β	SE(β)	z	p
(INTERCEPT)	1.81	0.22	8.3	<**.0001**
AMBIG.VS.REST	−2.56	0.45	−5.6	<**.0001**
COMPLEX.VS.EASIER	−3.20	0.53	−6.0	<**.0001**
SIMPLE.VS.UNAMBIG	−2.68	0.81	−3.3	<**.001**
TRIAL	0.00	0.01	0.3	0.8
TRIAL:AMBIG.VS.REST	−0.07	0.03	−2.6	<**.05**
TRIAL:COMPLEX.VS.EASIER	−0.01	0.03	−0.4	0.7
TRIAL:SIMPLE.VS.UNAMBIG	0.08	0.05	1.7	**0.08**

Table 2: Model output of Exp. 1. AMBIG.VS.REST, COMPLEX.VS.EASIER, and SIMPLE.VS.UNAMBIG are the Helmert-coded condition contrast predictors, in order.

dom slopes for CONDITION and TRIAL and by-item random intercepts. Results are given in Table 2.

All Helmert contrasts reached significance at $p < .001$. That is, all target/competitor distributions shown in Fig. 3(a) are different from each other. There was no main effect of TRIAL, indicating that no learning took place overall during the course of the experiment. However, there were significant interactions, suggesting selective learning in a subset of conditions. In particular there was a significant interaction between TRIAL and the Helmert contrast coding the difference between ambiguous fillers and the rest of the conditions (AMBIG.VS.REST, $\beta = -.05$, $SE = .02$, $p < .05$) and a marginally significant interaction between TRIAL and the Helmert contrast coding the difference between the simple implicature and unambiguous filler condition (SIMPLE.VS.UNAMBIG, $\beta = .08$, $SE = .05$, $p = .08$). Further probing the simple effects revealed that participants chose the target more frequently later in the experiment in the simple and complex condition. This was evidenced by a main effect of TRIAL on that subset of the data ($\beta = .03, SE = .01, p < .05$) but no interactions with condition. There were no learning effects in the ambiguous and unambiguous filler conditions; participants were at chance for ambiguous items and at ceiling for unambiguous items throughout. This suggests that at least some participants became aware that there was an optimal strategy and began to employ it as the experiment progressed.

We next address the question of whether the data supports the within-participant distributions predicted by standard IBR. Recall from Section 2 that for the simple condition, IBR predicts R_0 players to have a uniform distribution over target and competitor choices and $R_{\geq 1}$ players to choose only the target. For the complex condition, the uniform distribution is predicted for both R_0 and R_1 players, while only target choices are expected for $R_{\geq 2}$ players.

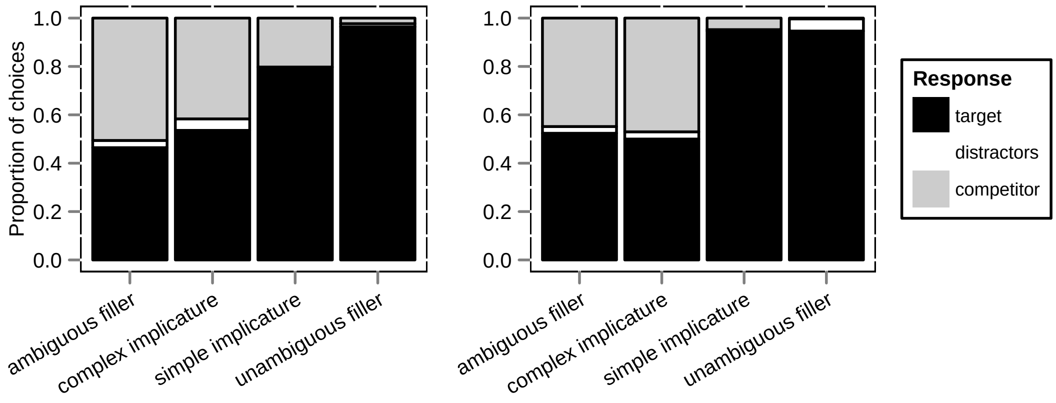
This is not borne out (see Fig. 4(a)). On the one hand, there were 3 participants in the simple condition and 5 in the complex condition who chose the target on half of the trials and could thus be classified as R_0 (or R_1 in the complex condition). Similarly, there were 11 participants in the simple condition and one in the complex condition who chose only targets and thus behaved as sophisticated receivers according to IBR. On the other hand, the majority of participants' distributions over target and competitor choices deviated from both the uniform and the target-only distribution.

One possibility is that some participants' type shifted from R_k to R_{k+1} as the experiment progressed. That is, they may have shifted from initially choosing targets and competitors at random to choosing only targets. However, while it is the case that overall more targets were chosen later in the experiment in both implicature conditions, there was nevertheless within-participant variation in choices late in the experiment inconsistent with a categorical shift. Another possibility is that the experiment was too short to observe this categorical shift.

5 Experiment 2

Exp. 2 tested participants' behavior in a *production* task that used instantiations of the signaling games described in Section 2.

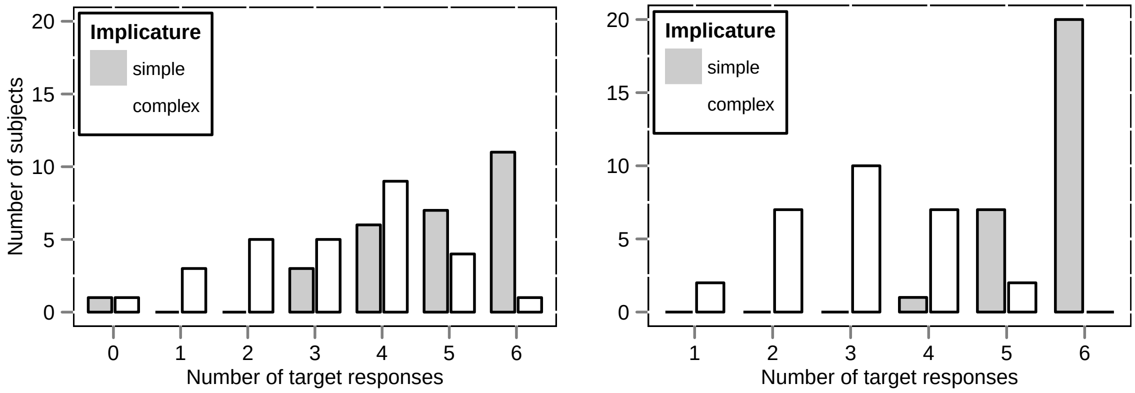
2009/05/14/random-effect-structure/



(a) Experiment 1

(b) Experiment 2

Figure 3: Proportions of target, competitor, and distractor choices in implicature and filler conditions (Exps. 1 & 2).



(a) Experiment 1

(b) Experiment 2

Figure 4: Distribution of participants over number of target choices in implicature conditions (Exp. 1 & 2).

5.1 Methods

Participants. Using Amazon's Mechanical Turk, 30 workers were paid $0.60 to participate under the same conditions as in Exp. 1. Data from two participants whose comments indicated that not all images displayed properly were excluded.

Procedure and Materials. The procedure was the same as on the sender trials in Exp. 1. Participants saw 36 trials with a 2:1 ratio of fillers to critical trials. There were 12 critical trials (6 simple and 6 complex implicature situations as in Fig. 1). Half of the fillers used the same displays as the implicature trials, but one of the other two objects was highlighted. This meant that the target message was either unambiguous (e.g. when the highlighted object was t_{t} in Fig. 1(a) the target message was m_{c}) or entirely ambiguous. The remaining 12 filler trials employed other displays with either entirely unambiguous or ambiguous target messages. Two experimental lists were created and counterbalancing was ensured as in Exp. 1.

5.2 Results and Discussion

Proportions of choice types are displayed in Fig. 3(b). As in Exp. 1, participants were close to ceiling for target message choices on unambiguous filler trials but at chance on ambiguous ones. On critical implicature trials, participants' performance was slightly different than in Exp. 1. Most notably, the distribution over target and competitor choices in the simple implicature condition was more skewed than in Exp. 1 (95% targets, 5% competitors), while it was more uniform than in Exp. 1 on complex implicature trials (50% targets, 47% competitors).

We again fitted a logistic mixed-effects regression model to the data. Trials on which the distractor messages were selected were excluded to allow for a binary outcome variable (target vs. competitor choice). This led to an exclusion of 2% of trials. In addition, the unambiguous filler condition is not included in the analysis reported here since there was only 1 non-target choice after exclusion of distractor choices, leading to unreliable model convergence. Thus, as in Exp. 1, CONDITION was entered into the model as a Helmert-coded variable but with only two contrasts, one comparing the simple implicature condition to the mean of ambiguous fillers and the complex implicature condition (SIMPLE.VS.HARDER), and another one comparing the ambiguous fillers with the complex implicatures (AMBIG.VS.COMPLEX). The model reported here further does not contain a TRIAL predictor to control for learning effects because model comparison revealed that it was not justified ($\chi^2(1) = 0.06$, $p = .8$). That is, there were no measurable learning effects in this experiment. We included the maximal random effects structure that allowed the model to converge: by-participant random slopes for CONDITION and by-item random intercepts.

The SIMPLE.VS.HARDER Helmert contrast reached significance ($\beta = 3.04$, $SE = 0.5$, $p < .0001$) while AMBIG.VS.COMPLEX did not ($\beta = 0.08$, $SE = 0.41$, $p = .9$). That is, there was no difference between choosing a target in the ambiguous filler condition and in the complex implicature condition, suggesting that participants were at chance in deriving complex implicatures in production. However, they were close to ceiling in choosing targets in the simple implicature condition.

The observed within-participant distributions are better predicted by the qualitative version of IBR than in Exp. 1 (see Fig. 4(b)). For the simple condition, IBR predicts S_0 players to have a uniform distribution over target and competitor choices and $S_{\geq 1}$ players to choose only the target. For the complex condition, the uniform distribution is predicted for both S_0 and S_1 players, while only target choices are expected for $S_{\geq 2}$ players.

In the simple implicature condition, 75% of participants were perfect S_1 reasoners. The remaining 25% chose almost only targets. That is, participants very consistently computed the implicature. In contrast, the bulk of participants chose targets versus competitors at random in the complex implicature condition. Only 2 participants chose the target 5 out of 6 times.

Comparing these results to the results from Exp. 1, we see the following pattern: in production the simple one-level implicatures are more readily computed than in comprehension, while the more complex two-level implicatures are more readily computed in comprehension than in production. That is, rather than comprehension mirroring production, in this paradigm there is an asymmetry between the two. This is consistent with the quantitative interpretation of IBR (as described in section 3) that takes into account players' uncertainty about communicative success.

6 General Discussion

In two studies using an abstract language game we investigated speakers' and hearers' strategic reasoning about referential descriptions. Most generally, our results clearly favor step-wise solution concepts like IBR over equilibrium-based solution concepts (e.g. Parikh (2001)) as predictors of participants' pragmatic reasoning: our results suggest that interlocutors do take perspective and simulate each others' beliefs, although (a) message and interpretation choice behavior is not always optimal and (b) perspective-taking decreases as the number of reasoning steps required to arrive at the optimal response, as predicted by IBR, increases.

We also found evidence for an intriguing asymmetry between production and comprehension. While not predicted by the standard formulation of the IBR model, this asymmetry is consistent with an interpretation of IBR that takes into account the uncertainty that interlocutors have about the probability of communicative success given a restricted set of message and interpretation options. This calls for a revision of the IBR model to incorporate more nuanced quantitative information. Since, moreover, there is a substantial amount of individual variation, further investigating the role of individual differences on perspective-taking (e.g. Brown-Schmidt (2009)) promises to be a fruitful avenue of further research that could inform model revisions.

It could be objected that the comparison of implicatures across experiments may be problematic due to the different nature of the tasks involved in the production vs. comprehension experiments and differences underlying the involved inference processes. However, note that the version of the IBR model that takes into account interlocutor uncertainty predicts the asymmetry between production and comprehension that we found precisely by integrating some of the differences involved in the two processes: most importantly, since conversation is modelled as a dynamic game, the sender reasons about the future behavior of the receiver, while the receiver reasons "backward", so to speak, using Bayesian conditionalization, about the most likely initial state the sender could have been in; this gives rise, as we have seen, to different predictions about when a speaker or a hearer can be absolutely certain of communicative success. How this difference is implemented mechanistically is an interesting question that merits further investigation.

Frank and Goodman (2012) report the results of an experiment using a referential game almost identical to ours and show that a particular Bayesian choice model very reliably predicts the observed data for both comprehension and production. In fact, the proposed Bayesian model is a variant of IBR reasoning that considers only a level-1 sender and a level-2 receiver, but assumes *smoothed* best response functions at each optimization step. In a smoothed IBR model, players' choices are stochastic with choice probabilities proportional to expected utilities (see Rogers et al. (2009) for a general formulation of such a model in game theoretic terms). This suggests a straightforward agenda for future work: combining our approach and that of Frank and Goodman (2012), smoothed IBR models that allow various strategic types for speakers and listeners should be further tested on empirical data.

In related work investigating comprehenders' capacity for deriving ad hoc scalar implicatures, Stiller et al. (2011) found that subjects could draw simple implicatures of the type we report above in a setup very similar to ours, but failed to draw complex ones. In contrast, our comprehenders performed above chance in the complex condition (albeit only slightly so). One possible explanation for this difference is that unlike Stiller et al. (2011), we restricted the set of message alternatives and also made it explicit to participants that a message could only denote one feature. This highlights the importance of (mutual knowledge of) the set of alternatives assumed by interlocutors in a particular communicative setting. While we restricted this set explicitly, in natural dialogue there is likely a variety of factors that determine what constitutes an alternative.

This suggests that future extensions of this work should move towards an artificial language paradigm. For example, whether a given message constitutes an alternative is likely to be affected by message complexity, which was held constant in our setup by using pictorial messages. Artificial language paradigms allow for investigating the effect of message complexity on inferences of the type reported here. Similarly, it will be important to further test the quantitative predictions made by IBR, e.g. by parametrically varying the payoff of communicative success and failure s and f and the interaction thereof with message complexity.

One question that arises in connection with the restrictions we imposed on the set of available pictorial messages, is the extent to which our results are transferable to natural language use. This is a legitimate concern that we would have to address empirically in future work. But notice also that, firstly, there is no *a priori* reason to believe that reasoning about natural language use and reasoning about our abstract referential games should necessarily differ — indeed it has been noted as early as Grice (1975) that conversational exchanges constitute but one case of rational communicative behavior. More importantly, even if reasoning about natural language *were* different in kind from strategic reasoning in general, the kind of strategic IBR reasoning we address here is a specific variety of reasoning that has been explicitly proposed in the literature as a model of pragmatic reasoning. The reported experiments are thus relevant in at least as far as they are the first empirical test of whether human reasoners are, in general, able to perform *this* kind of strategic reasoning in a task that translates the proposed pragmatic context models as directly as possible into an experimental setting.

We conclude that the studies reported are an encouraging first step towards validating game-theoretic approaches to formal pragmatics, which are well-suited to modeling pragmatic phenomena and generating quantitative, testable predictions about language use. The future challenge, as we see it, lies in fine-tuning the formal models alongside further careful empirical investigation.

Acknowledgements

We thank Gerhard Jäger, T. Florian Jaeger, and Michael K. Tanenhaus for fruitful discussion. This work was partially supported by a EURO-XPRAG grant awarded to the authors and NIH grant HD-27206 to Michael K. Tanenhaus.

References

Anton Benz and Robert van Rooij. 2007. Optimal assertions and what they implicate. *Topoi*, 26:63–78.

Sarah Brown-Schmidt, Christine Gunlogson, and Michael K. Tanenhaus. 2008. Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition*, 107:1122–1134.

Sarah Brown-Schmidt. 2009. The role of executive function in perspective taking during online language comprehension. *Psychonomic Bulletin and Review*, 16(5):893 – 900.

Vincent P. Crawford and Nagore Iriberri. 2007. Fatal attraction: Salience, naïveté, and sophistication in experimental ”hide-and-seek” games. *The American Economic Review*, 97(5):1731–1750.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998.

Michael Franke. 2011. Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics & Pragmatics*, 4(1):1–82.

Dave Golland, Percy Liang, and Dan Klein. 2010. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 410–419, Cambridge, MA. Association for Computational Linguistics.

H.P. Grice. 1975. Logic and conversation. *Syntax and Semantics*, 3:41–58.

Daniel Grodner and Julie C. Sedivy. 2011. The effect of speaker-specific information on pragmatic inferences. In N. Pearlmutter and E. Gibson, editors, *The Processing and Acquisition of Reference*. MIT Press, Cambridge, MA.

Daniel Grodner, Natalie M. Klein, Kathleen M. Carbary, and Michael K. Tanenhaus. 2010. ”Some”, and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition*, 116:42 – 55.

Joy Hanna, Michael K. Tanenhaus, and John C. Trueswell. 2003. The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, 49:43–61.

Trey Hedden and Jun Zhang. 2002. What do you think i think you think?: Strategic reasoning in matrix games. *Cognition*, 85(1):1–36.

Daphna Heller, Daniel Grodner, and Michael K. Tanenhaus. 2008. The role of perspective in identifying domains of reference. *Cognition*, 108:831–836.

Y. Huang and Jesse Snedeker. 2009. On-line interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cognitive Psychology*, 58:376–415.

Gerhard Jäger. 2008. Applications of game theory in linguistics. *Language and Linguistics Compass*, 2/3:406–421.

Gerhard Jäger. 2011. Game-theoretical pragmatics. In Johan van Benthem and Alice ter Meulen, editors, *Handbook of Logic and Language*, pages 467–491. Elsevier, Amsterdam.

Boaz Keysar, Dale J. Barr, and J. S. Brauner. 2000. Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, 11:32–37.

Boaz Keysar, S. Lin, and Dale J. Barr. 2003. Limits on theory of mind use in adults. *Cognition*, 89:25–41.

Prashant Parikh. 2001. *The Use of Language*. CSLI Publications, Stanford University.

Brian W. Rogers, Thomas R. Palfrey, and Colin Camerer. 2009. Heterogeneous quantal response equilibrium and cognitive hierarchies. *Journal of Economic Theory*, 144(4):1440–1467.

Julie C. Sedivy. 2003. Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, 32:3–23.

Alex Stiller, Noah D. Goodman, and Michael C. Frank. 2011. Ad-hoc scalar implicature in adults and children. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*.