Rationalizable signaling

Gerhard Jäger*

University of Tübingen
and
Swedish Collegium for Advanced Study, Uppsala

September 2012

Sage nicht alles, was du weißt, aber
wisse immer, was du sagst.
Don't say everything that you know, but always know what youre saying.

Matthias Claudius (1740 – 1815)

Abstract

An important finding of the game theoretic research on signaling games is the insight that under many circumstances, a signal obtains credibility by incurring costs to the sender. Therefore it seems questionable whether or not *cheap talk* — signals that are not payoff relevant — can serve to transmit information among rational agents. This issue is non-trivial in strategic interactions where the preferences of the players are not aligned.

Researchers like Crawford & Sobel, Rabin, and Farrell demonstrated, however, that even in the case of partially divergent interests, cheap talk may be informative. They assume that signals have an exogenously given meaning that is common knowledge between the players, and they explore the conditions under which such a signal is credible.

This discussion has obvious relevance to the program of Gricean pragmatics in linguistics. According to Grice's "Cooperative Principle", this research tradition only considers scenarios where the interests of the players are aligned. Nevertheless the assumption of differential signaling costs introduces an element of non-aligned interests here.

The present paper proposes a framework that combines these two research strands. Using an inference protocol of *iterated best response*, it gives a recipe how the interlocutors derive rationalizable strategies from exogeneously given "literal" meanings of signals. No special assumptions about the alignment of interests or signaling costs are made.

After introducing the formal model, the paper sketches several applications of this models to problems in linguistic pragmatics, like scalar implicatures, the division of pragmatic labor, and the interpretation of measure terms.

*Address for correspondence: *University of Tübingen, Department of Linguistics, Wilhelmstr. 19, 72072 Tübingen, Germany, phone: +49-7071-2977302, fax: +49-7071-295213, email: gerhard.jaeger@uni-tuebingen.de*

1 Introduction

On August 16, 2012, the Republican presidential candidate Mitt Romney made the following statement about the taxes he paid over the preceding years:

> "But I did go back and look at my taxes and over the past 10 years I never paid less than 13 percent. I think the most recent year is 13.6 or something like that."

If Romney paid 20 percent in taxes in all these years except the previous one, this statement would still be true. Also, it might be possible that the statement is untrue, and he in fact only paid 10 percent during all those years. Nevertheless, most pundits will infer from this statement that Romney's minimum tax rate over the past 10 years is between 13 and 14 percent, that his average tax rate in the same period is not much higher, and that in 2011, he paid between 13.5 and 13.7 percent (probably closer to 13.5 than to 13.7).

These inferences are based (a) on common knowledge about Romney's preferences and (b) the assumption that he is rational. It is certainly in his interest to say the truth. Otherwise some newspaper or blog might find out that he was lying, and this would severely damage his presidential ambitions. Furthermore, he probably prefers voters to believe that he paid a high tax rate. So if his average tax rate had been below 13 percent, he wouldn't have dared to make this statement, and if it had been above 13 percent, he would have named a higher figure. Regarding the most recent year, he chose a rather precise figure with one digit after the decimal point, plus the qualifying expression *or something like that*. This indicates that there is a certain margin or error, which is in the order of magnitude of 0.1 percent.

This example illustrates several important points about communication between rational agents using a common language. It shows that the truth conditions of a message are crucial for its interpretation. However, the information that is transmitted is usually not identical with these truth conditions. Rather, it is derived from them via rational deliberation, taking the strategic preferences of sender and addressee(s) into account.

The tension between what is said and what is meant, and the systematic relationship between the two, is the central concern of linguistic pragmatics in the tradition of Paul Grice's work (see especially Grice 1975). Grice assumes that conversation is a kind of rational interaction, and therefore general principles of rational interaction apply. Grice's famous Maxims of Conversation, being subordinate to the Cooperation Principle, give an informal description of the preferences of the interlocutors. They serve as heuristics to compute what is meant from what is said.

Game theory is a branch of applied mathematics that specifically deals with the principles of interaction among rational agents. It is thus natural to employ game theory for a formalization of the Gricean program. To put in in the words of Robert Stalnaker (2005):

> "As many people have noticed, Gricean ideas naturally suggest a game theoretic treatment. The patterns of iterated knowledge and belief that are characteristic of game theoretic reasoning are prominent in Grice's discussions of speaker meaning, and the pattern of strategic reasoning that Grice discussed in the derivation of conversational implicatures are patterns that game theory is designed to clarify. ... [G]ame theory provides some sharp tools for formulating some of Grice's ideas[.] ... And I think Gricean ideas will throw some light on the problems game theorists face when they try to model communicative success."

It is worth noting that information transmission among rational agents that have a common language (i.e. a language where signals have a literal meaning that is common knowledge between the interlocutors) has been investigated in the economics literature as well. Key publications from this area Rabin (1990) and Farrell (1993). There, and in subsequent work, it is usually assumed that the preferences of the interlocutors are not identical. A central question is under what conditions genuine information transmission is possible.

Game theoretic approaches to linguistic pragmatics has become an active area of research in recent years. There are essentially two approaches that are being pursued. Prashant Parikh (see for instance Parikh 2001, 2010) assumes that rational communication takes place in a *Nash equilibrium*, i.e. that each agent behaves rationally given the behavioral disposition of the other agent. This is a standard assumption in most work in game theory. If a game has more than one equilibrium, the players have to agree on a protocol for equilibrium selection to achieve efficient interaction. Parikh gives a such a criterion for equilibrium selection, so-called Pareto optimality. Similarly, Robert van Rooij (see for instance van Rooij 2004) proposes that pragmatic reasoning is based on selecting a specific Nash equilibrium, namely an evolutionarily stable one.

In Parikh's model, it is part of the structure of the game that the sender necessarily sends a true message. If this assumption is lifted, uniqueness of a Pareto optimal Nash equilibrium is not guaranteed. Quite generally, if there is no *a priori* preference for truthfulness, communication games have many symmetric equilibria, and it is not possible to select among them solely on the basis of rationality considerations.[1]

This is one of the reasons why the other research tradition of game theoretic pragmatics shuns the notion of Nash equilibria. Following the basic idea from Rabin (1990), it is rather assumed that truthfulness is a kind of default assumption that is overridden only if the rational self-interests of the interlocutors requires it. Rationality considerations may or may not lead to a Nash equilibrium. They will, however, always lead to strategies that are *rationalizable*, i.e. that are consistent with the assumption that it is common knowledge between the interlocutors that they are both rational.

In this paper, a detailed implementation of this rationalizability based research program is spelled out. In comparison to related proposals such as Franke (2009, 2011), I will make very weak assumptions about the knowledge states of the interlocutors. In particular, they may have only partial knowledge about each others preferences, and they may employ private knowledge that is not known to the other player.

In the next section, I will briefly recapitulate some basic ideas about rational communication from the game theoretic literature. I will modify this model by incorporating some concepts from Gricean pragmatics informally in Section 3, and I will present a formal framework for computing pragmatic interpretation from the literal meaning of expressions and the preferences and beliefs of the language users in Section 4. Section 5 contains a series of examples that serve to illustrate the empirical predictions of this model and to compare it to other neo-Gricean theories such as bidirectional Optimality Theory. In Section 6 this model is compared to Michael Franke's *IBR model* of game theoretic pragmatics. Sections 7 and 8 contain additional pointers to related work and concluding remarks.

2 Signaling games

Let us consider a very elementary example for a situation where communication may make a difference. Suppose Sally pays Robin a visit, and Robin wants to offer his guest something to drink, either tea or coffee. Sally is either a tea drinker or a coffee drinker, and Robin prefers the outcome where Sally receives her favorite drink over the other outcome, but he does not know Sally's preferences.

We may formalize this scenario as follows: There are two possible worlds, w_1 and w_2. In w_1, Sally prefers tea; in w_2 she prefers coffee. Robin has a choice between two actions. Offering tea would be action a_1, and offering coffee is action a_2. Sally knows which world they are in, but Robin does not know it. Let us assume that Robin assigns both worlds an *a priori* probability of 50%. So the scenario can be represented by Table 1. Rows represent possible worlds and columns represent Robin's actions. The first number in each cell gives Sally's payoff for this configuration, and the second number Robin's payoff.

	a_1	a_2
w_1	$1;1$	$0;0$
w_2	$0;0$	$1;1$

Table 1: A simple coordination scenario

Without any further coordination between the players, Robin will receive an expected payoff of 0.5 for either action, and hence Sally will also receive, on average, a payoff of 0.5. They can do better though if they communicate. Suppose it Robin expects that Sally says "tea" in w_1 and "coffee" in w_2. Then the rational course of action for Robin is to perform a_1 if he hears "tea", and to perform a_2 upon hearing "coffee". If Sally knows that Robin will react to these signals in this way, it is in fact rational for her to say "tea" in w_1 and "coffee" in w_2.

So adding the option for communication may improve the payoff of both players. Technically, the original scenario (which is not really a game but a decision problem because Sally has no choice to make) is transformed into a *signaling game*. Here the sender (Sally in the example) can send signals, and she can condition the choice of signals on the actual world. So a strategy for the sender is a function from possible worlds to signals. The receiver (Robin in the example) can condition his action on the signal received. So a strategy for the receiver is a function from signals to actions. (Analogous scenario's are studied extensively by Lewis 1969.)

The above example suggests that rational players will benefit from the option of communication. Things are not that simple though. If Sally says "I want tea" in w_1 and "I want coffee" in w_2, and Robin interprets "I want tea" as a_1 and "I want coffee" as a_2, both players benefit. Let us call this mapping from world to signals to actions L_1. They would receive the same benefit though if Sally said "I want coffee" in w_1 and "I want tea" in w_2, and Robin interprets "I want coffee" as a_1 and "I want tea" as a_2, which I will call L_2.[2] Pure reason does not provide a clue to decide between these two ways to coordinate. It is thus consistent with rationality that Sally assumes Robin to use L_2 and thus to signal according to L_2, while Robin assumes Sally to use L_1, and thus will interpret her signals according to L_1. In this situation, Robin will perform a_2 in w_1 and a_1 in w_2. Both players would receive the worst

possible expected payoff of 0 here.

These considerations ignore the fact that the two signals do have a conventional meaning which is known to both players. L_1 is *a priori* much more plausible than L_2 because in L_1 Sally always says the truth, and Robin always believes the literal meaning of Sally's message.

Rational players cannot always rely on the honesty/credulity of the other player though. Consider the scenario from Table 2 (unless otherwise indicated, I will always assume a uniform probability distribution over possible worlds):

	a_1	a_2
w_1	$1;-1$	$-1;1$
w_2	$-1;1$	$1;-1$

Table 2: A simple zero sum scenario

Here the interests of Sally and Robin are strictly opposed; everybody can only win as much as the other one looses. Here too, there are two signals that Sally can send. We call them f_1 and f_2. They both have a conventional literal meaning: $[\![f_1]\!] = \{w_1\}$ and $[\![f_2]\!] = \{w_2\}$. If Robin is credulous, he will react to f_1 with a_2 and to f_2 with a_1. If Sally believes this and is rational, she will be dishonest and send f_1 in w_2 and f_2 in w_1. If Robin is not quite so credulous, he may anticipate this and switch his strategy accordingly, etc. In fact, it turns out that with or without communication, any strategy is rationalizable in this game.[3] The lesson here is that communication might help in situations where the interests of the players are aligned, but it does not make a difference if these interests are opposed.

The most interesting scenarios, of course, are those where the interests of the players are partially, but not completely aligned. In a very influential paper, Crawford and Sobel (1982) showed that in such intermediate scenarios communication can be beneficial to both players. I will discuss a modified version of their scenario here.

Suppose Sally is starting in a new job, and Robin is her employer. Sally has a certain skill level that can be expressed by some (possibly negative) integer w. Robin will employ her at a certain level in the professional hierarchy that is summarized by some real number a.[4] According to Robin's preferences, Sally's position in the hierarchy is as close as possible to her skill level. Let us say that his utility function is

$$u^R(w, a) = -(w - a)^2$$

Sally would prefer to be placed somewhat higher in the hierarchy, because this would give her a higher income. However, she does not want to be over-estimated too much because the strain might be too hard. Ideally, she wants to be over-estimated by 1. So her utility function is

$$u^S(w, a) = -(w + 1 - a)^2$$

We assume that each integer is a possible signal, and a signal f is true of some skill level w iff $f = w$.

When filling out her paperwork, Sally has to specify her skill level. Suppose Sally uses some signal f. If Robin trusts her, he will choose $a = f$. Sally has an incentive to exploit this and to choose the signal $f = w + 1$, where w is her actual skill level. If Robin anticipates this, he will choose the action $a = f - 1$ instead. Taking this into account, Sally should

actually choose the signal $f = w + 2$. If Robin expects this, he will choose $a = f - 2$, which incites Sally to use $f = w + 3$ etc. As long as there are no penalties for exaggeration and it is not specified how far Sally is prepared to go in her deception, it is not possible to draw any conclusions beyond $w \leq f$ from Sally's signal.

Now suppose there are only two messages, f_+ and f_-. The literal meaning of f_+ and f_- are $w \geq -2$ and $w \leq -3$ respectively. Furthermore, let us assume that Robin's prior probability function over Sally's skill level, call it p^*, is as follows:

$$p^*(w) = \frac{2^{-|w|}}{3}.$$

If Robin is credulous, upon observing f_+ he will choose the action that maximizes his expected utility given p^* and f_+'s truth conditions. This is

$$\begin{aligned} \arg\max_a \sum_{w \in [[f_+]]} u^R(w,a) p^*(w | w \in [\![f_+]\!]) &= \sum_{w=-2}^{\infty} w p^*(w) / \sum_{w=-2}^{\infty} p^*(w) \\ &= 4/11 \approx 0.36.^5 \end{aligned}$$

Conversely, upon observing f_- and believing it, Robin's best choice is

$$\begin{aligned} \arg\max_a \sum_{w \in [[f_-]]} u^R(w,a) p^*(w | w \in [\![f_-]\!]) &= \sum_{w=-\infty}^{-3} w p^*(w) / \sum_{w=-\infty}^{-3} p^*(w) \\ &= -4. \end{aligned}$$

If $w \leq -3$, i.e. if f_- is true and if Sally expects Robin to believe her message, it is in Sally's interest to use f_- rather than f_+, because then -4 is closer to her desired action $w+1$ than $4/11$. On the other hand, if $w \geq -2$, f_+ is the better choice. So in each w, it is rational for Sally to say the truth. Symmetrically, if Robin expects Sally to say the truth, it is rational for him to believe her. This pair of strategies — the honest sender strategy and the credulous receiver strategy — actually form a Nash equilibrium. This means that no player would have an incentive to change their strategy if they knew the other player's strategy.

This example illustrates two important insights: (a) it can be rational for both players to use communication even if their interests are not completely aligned, and (b) whether or not it is rational to be honest/credulous may depend on the space of available messages. If Sally could use signals with a very specific meaning, this might tempt her into trying to deceive Robin, which, if anticipated, would lead to a breakdown of communication. If only sufficiently vague messages are available, this temptation does not arise.

Table 3 represents another example that may illustrate this point. It is taken from Rabin (1990).

In w_1, Sally's and Robin's interests are identical; they both want Robin to take action a_1. So if Sally sends f_1, Robin has no reason to doubt it, and he will react to it by performing a_1. *Prima facie*, it might seem that the same holds for w_2. Here, both players would prefer Robin to take action a_2. However, in w_3 Sally also prefers Robin to take action a_2, while Robin would prefer a_3 if he knew that w_3 is the case. So in w_3 the interests of the players diverge, and Sally might be tempted to send the signal f_2 both in w_2 and w_3. Robin is thus well-advised not to believe f_2 in its literal meaning. If he does not know whether he is in w_2

	a_1	a_2	a_3
w_1	$10;10$	$0;0$	$0;0$
w_2	$0;0$	$10;10$	$5;7$
w_3	$0;0$	$10;0$	$5;7$

Table 3: Partially aligned interests

or in w_3, his rational action is to hedge his bets and to perform a_3 after all, which guarantees him an expected payoff of 7 (against an expected payoff of 5 for a_2).

After performing these reasoning steps, Sally will perhaps convince herself that she has no chance to manipulate Robin into performing a_2. The best thing she can do both in w_2 and in w_3 is to prevent him from performing a_1. It is thus rational for Sally to say f_{23} (where $[\![f_{23}]\!] = \{w_2, w_3\}$) both in w_2 and w_3. So the situation that is most beneficial for both players is the one where only the signals f_1 and f_{23} are used, Sally uses each signal if and only if it is true, and Robin believes her and acts accordingly. In Rabin's terminology, f_1 is *credible* in this scenario, while f_2 would not be credible. Simplifying somewhat, a message f is credible (according to Rabin 1990) if it is rational for the sender to use it whenever it is true provided she can expect it to be believed, and if it is rational for the receiver to act as if she believed it provided the sender uses it whenever it is true.

Incidentally, f_{23} would not come out as credible in Rabin's model. The reason is that in w_2 or w_3, Sally might try to convince Robin that they are in w_2 and thus to induce action a_2. This is not possible via credible communication, but Sally might believe that she is capable to outsmart Robin by taking some other action.[6]

As was argued in the beginning, rationality alone is insufficient to coordinate players in such a way that signals receive a stable interpretation. This is even the case if signals do have a conventionalized meaning that is known to all players (as is the case for expressions from some natural language if both players know that language). Rabin proposes that, beyond being rational, reasonable sender will always send a true credible message if this is possible, and reasonable receivers will always believe any credible message.[7] In many cases, this reduces the space of rationalizable strategies significantly and thus ensures a certain amount of information transmission that is in the interest of both players.

3 Gricean reasoning

The kind of reasoning that was informally employed in the last section is reminiscent to pragmatic reasoning in the tradition of Grice (1975). First, information can only be exchanged between rational agents if it is in the good interest of both agents that this information transfer takes place. This intuition, which Grice captured in his Cooperative Principle, is implicit in the notion of credibility. Also, Rabin adopts a default assumption that messages are used according to their conventional meaning, unless overarching rationality considerations dictate otherwise. This corresponds to Grice's Maxim of Quality. Furthermore, the first part of the Maxim of Quantity — "make your contribution as informative as is required (for the current purpose of the exchange)" — is implicit in the notion of rationality. For instance, suppose Sally is in world w_1 in the scenario described in Table 3. Then rationality requires her to transmit the information $\{w_1\}$ if there is a reliable way of doing so. If she would send a

	a_1	a_2	a_3
w_1	10; 10	0; 0	9; 9
w_2	0; 0	10; 10	9; 9

	a_1	a_2	a_3
w_1	10; 10	0; 0	6; 6
w_2	0; 0	10; 10	6; 6

Table 4: context c_1: scalar implicature/ context c_2: no scalar implicature

message which Robin would interpret as $\{w_1, w_2\}$, this would leave Robin in a state where he does not know whether a_1 or a_2 is the appropriate action. So sending an under-informative message would be irrational for Sally.

Despite these similarities, there are some crucial differences between Rabin's model and Gricean reasoning. To illustrate this, let us consider a schematic example of a scalar implicature. The utility structure is given in the left panel of Table 4. Suppose, as before, that there are three messages, f_1, f_2 and f_{12}, with the conventionalized meanings $\{w_1\}$, $\{w_2\}$, and $\{w_1, w_2\}$ respectively. However, we now assume that sending a message may incur some costs for the sender, and that different messages incur different costs. In the specific example, we assume that $c(f_1) = c(f_{12}) = 0$, and $c(f_2) = 2$, where $c(f)$ is the cost that the sender has to pay for sending message f. So the sender's utility is now a three-place function u^S that depends on the actual world, the message sent, and the action that the receiver takes. If $v^S(w, a)$ is the distribution of sender payoffs that is given in the left panel of Table 4 above, the sender's overall utility is

$$u^S(w, f, a) = v^S(w, a) - c(f)$$

You can imagine that Robin wants to know who was at the party last night, and Sally knows the answer. In w_1, all girls were at the party, and in w_2 some but not all girls were there. f_1 is the message "All girls were at the party", f_2 is "Some but not all girls were at the party", and f_{12} is "Some girls were at the party." Obviously f_2 is more complex than the other two messages, which are approximately equally complex. This is covered by the assignment of costs.

According to Gricean pragmatics, Sally would reason roughly as follows: If I am in w_1, I want Robin to perform a_1 because this gives me a utility of 10. a_1 is what he would do if he believed that he is in w_1. I can try to convince him of this fact by saying f_1. It is not advisable to say f_2, because if Robin believed it, he would perform a_2, which gives me a utility of a mere -2. Also saying f_{12} is not optimal because if Robin believes it, he will perform a_3, leading to a utility of 9. So it seems reasonable to send f_1 in w_1.

If we are in w_2, it might seem reasonable to say f_2 because if Robin believes it, he will perform a_2, which is my favorite outcome. However, I will have to pay the costs of 2, so my net utility is only 8. If I say f_{12} and Robin believes it, he will perform a_3. As f_{12} is costless for me, my net utility is 9, which is better than 8. So in w_2 I will send f_{12}.

Robin in turn will anticipate that Sally will reason this way. If he is confronted with the message f_1, he will infer that he is in w_1, and he will perform a_1. If he hears f_{12}, he will infer that w_2 is the case, and he will perform a_2 after all.

Sally, being aware of this fact, will reason: This taken into consideration, it is even more beneficial for me to send f_{12} if I am in w_2 because this will give me the maximal payoff of 10. So I have no reason to change the plan of sending f_1 in w_1 and f_{12} in w_2.

This reasoning leads to a sender strategy where f_{12} is sent if and only if $\{w_2\}$ is true. Following Lewis (1969), we will call the set of worlds where a certain message is sent its *indicative meaning* (as opposed to its imperative meaning, which is the set of actions that the receiver might perform upon receiving that message). In our example, the indicative meaning of f_{12} thus turns out to be $\{w_2\}$, which is a proper subset of its literal meaning $\{w_1, w_2\}$. The information that w_1 is not the case is a scalar implicature — "some" is pragmatically interpreted as "some but not all."

The reasoning pattern that is used here makes implicit use of the notion of the *best response* of a player to a certain probabilistic belief. A best response to a belief state is a strategy that maximizes the expected payoff of the player as compared to all strategies at their disposal, given this belief state. Rational players will always play some best response to their beliefs.

Suppose an external observer (i.e. Robin, or Sally trying to figure out Robin's expectations, or Robin trying to figure out Sally's expectations about his intentions etc., or we as modelers) has some partial knowledge about Sally's belief state. There is some set of receiver strategies R, and the observer knows that Sally expects Robin to play some strategy from R, and that Sally cannot exclude any element of R for sure. The observer does not know which probability Sally assigns to the elements of R. Then any probability distribution over R that assigns positive probabilities to all elements of R is a possible belief state of Sally's, as far as the observer's knowledge is concerned. Hence any best response of Sally's to such a belief state is a potential best response, or **cautious response**, as Pearce 1984 calls it, of Sally's to R. All the observer can predict with certainty if he assumes Sally to be rational is that she will play some cautious response to R. (Since Sally holds a specific private belief, she will actually only consider a subset of the cautious responses, but the observer does not know which one.)

The iterative inference process that was used in the computation of the implicature above can be informally described as follows:

- Sally provisionally assumes that Robin is entirely credulous, and that he conditions his actions only on the literal interpretation of the message received. Let us call the set[8] of credulous strategies R_0. In the first round of reasoning, Sally might ponder any strategy that is a cautious response to R_0. Let us call this set of strategies S_0.
- In the next round, Robin might ponder all strategies that are cautious responses to S_0. The set of these strategies is R_1.
- ...
- S_n (R_{n+1}) is the set of strategies that are cautious responses to R_n (S_n).
- If a certain strategy S (R) cannot be excluded by this kind of reasoning (i.e. if there are infinitely many indices i such that $S \in S_i$ ($R \in R_i$)), then S (R) is a *pragmatically rationalizable strategy*.

In the example, the scalar implicature arises because the difference between $v^S(w_2, a_2)$ and $v^S(w_2, a_3)$ is smaller than the costs of sending f_2. Suppose the utilities would be as in the right panel of Table 4, rather than as in the left panel. Then the pragmatically rationalizable outcome would be that Sally uses f_2 in w_2, while f_{12} would never be used. Informally speaking, the reasoning here relies on a tension between the Maxim of Quantity

and the Maxim of Manner.[9] The implicature only arises if the utilities are such that Manner wins over Quantity.

In a more realistic scenario, Robin might actually not know for sure what Sally's precise preferences are. If we call the utility matrix in Table 4 *context* c_1[10], and the utilities in the right panel in Table 4 context c_2, Robin might hold some probabilistic belief about whether Sally is in c_1 or in c_2. Likewise, Sally need not know for sure which context Robin is in. Now in each round of the iterative reasoning process, the players will ponder each strategy that is a cautious response to any probability distribution over contexts and strategies in the previous round.[11]

Sally's reasoning will now start as follows: In w_1, I will definitely send f_1, no matter which context I am in. If I am in context c_1, it is better to send f_{12} if I am in w_2 because the costs of sending the more explicit message f_2 exceed the potential benefits. If I am in c_2 and w_2, however, it is advisable to use f_2.

Robin, in turn, will reason: If I hear f_1, we are definitely in w_1, and the best thing I can do is to perform a_1, no matter which context we are in. If I hear f_2, we are in c_2/w_2, and I will perform a_2. If I hear f_{12}, we are in c_1/w_2, and I will also play a_2.

So in S_1 Sally will infer: f_1 will induce a_1, and both f_2 and f_{12} will induce a_2, no matter which context Robin is in. Since f_{12} is less costly than f_2, I will always use f_1 in w_1 and f_{12} in w_2, regardless of the context I am in. Robin, in R_1, will thus conclude that his best response to f_1 is always a_1, and his best response to f_{12} is a_2. Nothing will change in later iterations. So here, the scalar implicature from "some" to "some but not all" will arise in all contexts, even though context c_2 by itself would not license it.

One might argue that this is not quite what happens in natural language use. Here we predict that f_2 would never be used. A more realistic outcome would be that f_2 is still interpreted as $\{w_2\}$, and that by using it, Sally conveys the message that it is very important to her that w_1 is in fact excluded.

What I believe is going on here is that there are also contexts where Sally does not know for sure which world she is in. In this case f_{12} might be sent in w_1 after all. Whether or not Robin derives the implicature in question would depend then on how much probability he assigns to this option.

To keep things simple, I will confine the technical model to be derived in this article to scenarios where the sender has complete factual knowledge, i.e. where she knows the identity of the actual world. A generalization to games where both players have incomplete information is worked out in Jäger (2011).

4 The formal model

In this section I will develop a formal model that captures the intuitive reasoning from the previous section.

A *semantic game* is a game between two players, the sender S and the receiver R. It is characterized by a set of contexts $\mathcal{C}$, a set of worlds $\mathcal{W}$, a set of signals $\mathcal{F}$, a set of actions $\mathcal{A}$, a probability distribution p^*, an interpretation function $[\![\cdot]\!]$, and a pair of utility functions u^S and u^R. In the context of this paper, I will confine the discussion to games where $\mathcal{C}$, $\mathcal{W}$, $\mathcal{F}$, and $\mathcal{A}$ are all finite. p^* is a probability distribution over $\mathcal{W}$. Intuitively, $p^*(w)$ is the *a priori* probability that the actual world is w. We assume that $p^*(w) > 0$ for all $w \in \mathcal{W}$.

A few words on the notion of a context, as it is used here, are in order. The fact that

Sally may be in one of several contexts reflects Robin's uncertainty about Sally's preferences. Technically, Sally's contexts could also be modeled as possible worlds. However, Robin's prior belief about the possible world Sally is in is common knowledge, while his prior beliefs about Sally's context is not. Therefore worlds and sender contexts are conceptually different, which justifies the technical distinction. Essentially, assuming different contexts for Sally amounts to saying that Sally's true utility function is some convex combination of the possible sender contexts.

Likewise, having different receiver contexts reflects Sally's incomplete information about Robin's preferences. Using receiver contexts therefore goes beyond the standard notion of signaling games, because in signaling games, the preferences of the receiver are assumed to be common knowledge.

Note that sender contexts and receiver contexts do not communicate with each other. Collapsing a sender context and a receiver context into a single structure (such as the two tables in Table 4) is merely a matter of notational convenience. If the receiver utilities in the two contexts in Table 4 where exchanged, we would still be dealing essentially with the same game.

$u^S \in \mathcal{C} \times \mathcal{W} \times \mathcal{F} \times \mathcal{A} \mapsto \mathbb{R}$ is the sender's utility function. There is some function $v^S \in \mathcal{C} \times \mathcal{W} \times \mathcal{A} \mapsto \mathbb{R}$ and some function $c \in \mathcal{F} \mapsto \mathbb{R}$ such that

$$u^S(c, w, f, a) = v^S(c, w, a) - c(f).$$

$u^R \in \mathcal{C} \times W \times \mathcal{A} \mapsto \mathbb{R}$ is the receiver's utility function. $[\![\cdot]\!] \in \mathcal{F} \mapsto \wp(W)$ is the semantic interpretation function that maps signals to propositions.

The space of pure sender strategies $\mathcal{S} = \mathcal{C} \times \mathcal{W} \mapsto \mathcal{F}$ is the set of functions from context/world pairs to signals. The space of pure receiver strategies $\mathcal{R} = \mathcal{C} \times \mathcal{F} \mapsto \mathcal{A}$ is the set of functions from context/signals pairs to actions.

The structure of the game is common knowledge between the players.

Some auxiliary notations: If M is a finite and non-empty set, $\Delta(M)$ is defined as

$$\Delta(M) = \{q \in (M \mapsto [0,1]) | \sum_{x \in M} q(x) = 1\}.$$

This is the set of probability distributions over M. A related notion is:

$$\text{int}(\Delta(M)) = \{q \in (M \mapsto (0,1]) | \sum_{x \in M} q(x) = 1\}.$$

This is the set of probability distributions over M where each element of M receives a positive probability. The difference is subtle but important. Both $\Delta(\cdot)$ and $\text{int}(\Delta(\cdot))$ can be used to model probabilistic beliefs. If we say that a player holds a belief from $\Delta(\mathcal{C})$, say, this means that they may exclude some contexts with absolute certainty. On the other hand, if Sally believes that Robin plays his strategy according to $\text{int}(\Delta(R))$ for some set $R \subseteq \mathcal{R}$, then Sally may have certain guesses, but she is not able to exclude any strategy from R with certainty. We will use this to capture the intuition that the players may have biases, but they do not have other sources of established beliefs about the intentions of the other players beyond the assumption that pragmatic rationality is common knowledge.

Definition 1 *Let $\phi \subseteq \mathcal{W}$ be a proposition and $p \in \Delta(\mathcal{W})$ be a probability distribution over worlds.*

$$A^*(c, \phi, p) \doteq \arg\max_{a \in \mathcal{A}} \sum_{w \in \phi} p(w) u^R(c, w, a)$$

$$A^*(c,\phi) \;\doteq\; A^*(c,\phi,p^*)$$

So $A^*(c,\phi,p)$ is the set of actions that might be optimal for the receiver if he is in context c, his (probabilistic) prior belief about the possible worlds is p, and this prior belief is updated with the information that he is in ϕ. $A^*(c,\phi)$ is the set of actions that the receiver believes to be optimal in c if he updates the prior belief p^* with ϕ. (Recall that p^* is Robin's prior probability distribution over $\mathcal{W}$.)

The central step in the iterative process described above is the computation of the set of strategies that maximize the expected payoff of a player against some probability distribution over contexts and strategies of the other player. The notion of a *best response* captures this.

Definition 2

- *Let $r^* \in \mathcal{R}$ be a receiver strategy, $\sigma \in \Delta(\mathcal{S})$ a probability distribution over $\mathcal{S}$, and $q \in \Delta(\mathcal{C})$ a probability over contexts. (σ, q) represent a belief of the receiver.*

  $$r^* \in BR(\sigma, q)$$

  (r^* is a best response *of the receiver to (σ,q)) iff*

  $$\forall c \in \mathcal{C} : r^* \in \arg\max_{r\in\mathcal{R}} \sum_{s\in\mathcal{S}} \sigma(s) \sum_{c'\in\mathcal{C}} q(c') \sum_{w\in\mathcal{W}} p^*(w) u^R(c,w,r(c,s(c',w)))$$

- *Let $s^* \in \mathcal{S}$ a sender strategy, $\rho \in \Delta(\mathcal{R})$ a probability distribution over $\mathcal{R}$, and $q \in \Delta(\mathcal{C})$ a probability over contexts. (ρ, q) represent a belief of the sender.*

  $$s^* \in BR(\rho, q)$$

  (s^* is a best response *of the sender to (ρ,q)) iff*

  $$\forall c \in \mathcal{C} \forall w \in \mathcal{W} : s^* \in \arg\max_{s\in\mathcal{S}} \sum_{r\in\mathcal{R}} \rho(r) \sum_{c'\in\mathcal{C}} q(c') u^S(c,w,s(c,w),r(c',s(c,w)))$$

In Pearce (1984), the notion of a **cautious response** against some set P of strategies is proposed. A cautious response to P is any pure strategy of the opposing player that is a best response to some probability distribution over P *that assigns positive probability to all members of P*.

Definition 3

- *Let $S \subseteq \mathcal{S}$ be a set of sender strategies. The set of* cautious responses to S *is defined as*

  $$CR(S) = \bigcup_{\sigma\in \mathrm{int}(\Delta(S))} \bigcup_{q\in\mathrm{int}(\Delta(\mathcal{C}))} BR(\sigma, q)$$

- *Let $R \subseteq \mathcal{R}$ be a set of receiver strategies. The set of* cautious responses to R *is defined as*

  $$CR(R) = \bigcup_{\rho\in \mathrm{int}(\Delta(R))} \bigcup_{q\in\Delta(\mathrm{int}(\mathcal{C}))} BR(\rho, q)$$

Suppose we know that Sally knows which context and world she is in, she believes for sure that Robin will play a strategy from R, and there is no more specific information that she believes to know for sure. We do not know which strategy from R Sally expects Robin to play with which likelihood, and which context Sally believes to be in. Under these conditions, all we can predict for sure is that Sally will play some strategy from $CR(R)$ if she is rational.

The same seems to hold if we only know that Robin expects Sally to play some strategy from S. Then we can infer that Robin, if he is rational, will certainly play a strategy from $CR(S)$. However, we may restrict his space of reasonable strategies even further. Suppose none of the strategies in S ever make use of the signal f. (Formally put, $f \in \mathcal{F} - \bigcup_{s \in S} \mathit{range}(s)$.) Then it does not make a difference how Robin would react to f, but he has to decide about the imperative meaning of f nevertheless (because receiver strategies are **total** functions from context/form pairs to actions). It seems reasonable to demand (and it leads to reasonable predictions, as we will see below) that Robin should, in the absence of evidence to the contrary, still assume that f is true. To take an example, suppose the teacher announces in class that whoever did the graffiti on his car will be punished. Then the teacher asks who did it. It is then rational for all students to deny it. If somebody raises their hand nevertheless, the teacher will be surprised but will assume that that person is guilty all the same.

If Robin encounters such an unexpected signal, he will have to revise his beliefs. In the previous paragraph I argued that this belief revision should result in an epistemic state where f is true. However, no further restrictions on Robin's belief revision policy will be stated. In particular, we will not demand that Robin will fall back to $p(\cdot|[\![f]\!])$, i.e. to the result of updating his prior belief with the literal interpretation of f. Robin will have to figure out an explanation why Sally used f despite his expectations to the contrary, and this explanation can bias his prior beliefs in any conceivable way. We have to assume though that the result of this believe revision is a consistent belief state, and that Robin will act rationally according to his new beliefs.

We can now proceed to define the iterative reasoning procedure that was informally described in the previous section.

Definition 4 (Iterated cautious response sequence)

$$\begin{aligned} R_0 &\doteq \{r \in \mathcal{R} | \forall c \in \mathcal{C} \forall f \in \mathcal{F} : r(c,f) \in A^*(c,[\![f]\!])\} \\ S_n &\doteq CR(R_n) \\ R_{n+1} &\doteq \{r \in CR(S_n) | \\ &\quad \forall f \in \mathcal{F} - \bigcup_{s \in S_n} \mathit{range}(s) \forall c \in \mathcal{C} \exists p \in \mathrm{int}(\Delta(\mathcal{W})) : r(c,f) \in A^*(c,[\![f]\!],p)\} \end{aligned}$$

This notion will be referred to as **ICR sequence** for short in the sequel.

R_0 is the set of credulous strategies of the receiver. S_n is the set of cautious responses of the sender against R_n. Likewise, R_{n+1} is the set of cautious responses of the receiver if he assumes that the sender plays a strategy from S_n in which he always tries to make sense of unexpected messages under the assumption that they are literally true.

The sets of *pragmatically rationalizable strategies* (PRS) are the set of sender strategies and receiver strategies that cannot be excluded for sure by the iterative reasoning process, no matter how deeply the reasoning goes.

This notion is related to Rabin's notion of message credibility. Intuitively, a message is credible iff it is used according to its literal meaning in all pragmatically rationalizable strategies. However, pragmatic rationalizability is more general as it also makes predictions about the usage of non-credible messages.

Definition 5 $(\mathbf{S}, \mathbf{R}) \in \wp(\mathcal{S}) \times \wp(\mathcal{R})$*, the sets of* pragmatically rationalizable strategies*, are defined as follows:*

$$\begin{aligned} \mathbf{S} &\doteq \{s \in \mathcal{S} | \forall n \in \mathbb{N} \exists m > n : s \in S_m\} \\ \mathbf{R} &\doteq \{r \in \mathcal{R} | \forall n \in \mathbb{N} \exists m > n : s \in R_m\} \end{aligned}$$

Note that there are only finitely many strategies in $\mathcal{S}$ and $\mathcal{R}$ (because we are only considering pure strategies). Therefore there are only finitely many subsets thereof. The step from (S_n, R_n) to (S_{n+1}, R_{n+1}) is always deterministic. It follows that the iterative procedure will enter a cycle at some point, i.e. there are n^* and i^* such that for all $m > n^*$ and for all k: $(S_m, R_m) = (S_{m+k \cdot i^*}, R_{m+k \cdot i^*})$. This ensures that $(\mathbf{S}, \mathbf{R})$ is always defined.

As was mentioned above, a strategy is called rationalizable iff a rational player might use it, provided it is common knowledge that all players are rational. The following formal definition (adapted from Osborne 2003:383) is provably equivalent to this informal characterization:

Definition 6 *The strategy pair* $(s^*, r^*) \in \mathcal{S} \times \mathcal{R}$ *is* rationalizable *iff there exist sets* $S \subseteq \mathcal{S}$ *and* $R \subseteq \mathcal{R}$ *such that*

- $S \subseteq \bigcup_{\rho \in \Delta(R)} \bigcup_{q \in \Delta(\mathcal{C})} BR(\rho, q)$
- $R \subseteq \bigcup_{\sigma \in \Delta(S)} \bigcup_{q \in \Delta(\mathcal{C})} BR(\sigma, q)$
- $(s^*, r^*) \in S \times R$

In words, s^* and r^* are rationalizable iff they are elements of some sets S and R such that every element of S is a best response to some belief of the sender that only considers strategies in R possible, and every element of R is a best response to some belief of the receiver that only considers strategies in S possible.

The notion of rationalizable strategies is closely related to the better-known notion of a Nash equilibrium. If two rational players have a way to coordinate their strategies (by pre-play communication or precedent, for instance), they will likely end up playing an equilibrium profile, i.e. a configuration of strategies each of which is rational given the strategies of the other player(s). If a game has several equilibria (such as all the games discussed here), rationality alone does not dictate the choice of a particular equilibrium. It can be said with certainty though that each player will play a rationalizable strategy — even in a one-shot game in the absence of any coordination devices — provided they are rational and it is common knowledge that all players are rational.

The set of pragmatically rationalizable strategies are in fact rationalizable:

Theorem 1 *For all* $(s^*, r^*) \in \mathbf{S} \times \mathbf{R}$*:* (s^*, r^*) *is rationalizable.*

Proof: As there are only finitely many subsets of $\mathcal{S}$ and $\mathcal{R}$ and (S_{n+1}, R_{n+1}) is a function of (S_n, R_n) for all n, there must be some $m^* \geq 0, i^* > 0$ such that for all $k, l \geq 0 : (S_{m^*+k \cdot i^*+l}, R_{m^*+k \cdot i^*+l}) = (S_{m^*}, R_{m^*})$. Let $s \in \mathbf{S}$. Then there must be some l^* such

that $s \in S_{m^*+l^*} = CR(R_{m^*+l^*})$ and $R_{m^*+l^*} \subseteq \mathbf{R}$. So there are $\rho \in \text{int}(\Delta(R_{m^*+l^*}))$ and $q \in \Delta(\mathcal{C})$ such that $s \in BR(\rho, q)$. Trivially, ρ can be extended to some $\rho' \in \Delta(\mathbf{R})$ by assigning zero probability to all elements of $\mathbf{R} - R_{m^*+l^*}$ such that $BR(\rho, q) = BR(\rho', q)$. So $s \in \bigcup_{\rho \in \Delta(\mathbf{R})} \bigcup_{q \in \Delta(\mathcal{C})} BR(\rho, q)$.

In a similar way, suppose $r \in \mathbf{R}$. Then there must be some l^* such that $r \in R_{m^*+l^*} = R_{m^*+i^*+l^*} \subseteq CR(S_{m^*+i^*+l^*-1})$ and $S_{m^*+i^*+l^*-1} \subseteq \mathbf{S}$. By an argument analogous to the previous case, it follows that $r \in \bigcup_{\sigma \in \Delta(\mathbf{S})} \bigcup_{q \in \Delta(\mathcal{C})} BR(\sigma, q)$. Hence

$$\mathbf{S} \subseteq \bigcup_{\rho \in \Delta(\mathbf{R})} \bigcup_{q \in \Delta(\mathcal{C})} BR(\rho, q)$$

and

$$\mathbf{R} \subseteq \bigcup_{\sigma \in \Delta(\mathbf{S})} \bigcup_{q \in \Delta(\mathcal{C})} BR(\sigma, q).$$

So any element of $\mathbf{S} \times \mathbf{R}$ is rationalizable. ⊣

This theorem is noteworthy because it demonstrates that players using the ICR recipe will end up behaving as if they are perfectly rational, even though ICR is a version of *bounded rationality*, involving only a limited reasoning depth and starting from default assumptions that do not presume rationality at all.

A note on computing the ICR sequence The epistemic notion of a cautious response is closely connected to the algorithmic notion of a *weakly undominated* (sometimes called *admissible*) strategy. Generally, in an n-person matrix game, strategy s for player i is weakly undominated iff there is no mixed strategies $x \in \Delta(S_i)$ of player i with

$$\begin{aligned} \forall t \in S_{-i}.u^i(s,t) &\leq \sum_{s' \in S_i} x_{s'} u^i(s',t), \text{ and} \\ \exists t \in S_{-i}.u^i(s,t) &< \sum_{s' \in S_i} x_{s'} u^i(s',t). \end{aligned}$$

(Following the standard notational convention in the literature, S_i denotes player i's strategy set, and S_{-i} the set of strategy combinations of all other players except i.) Strategy s is weakly undominated iff it is a cautious response to the set of strategy combinations from S_{-i} (see Pearce 1984, Appendix B for a proof).

To compute the cautious responses to a set R of receiver strategies, we determine the cautious responses to R for each speaker context c and possible world w separately. We construct a three-person game with the message, the receiver, and the receiver context as players. The utility for a triple (f, r, c') (with $s \in \mathcal{F}, r \in R$, and $c \in \mathcal{C}$) is defined as

$$U^{c,w}(f, r, c') = u^S(c, w, f, r(c', f)).$$

The cautious responses to R are those sender strategies that map each context-world pair (c, w) to some message f that is weakly undominated in $U^{c,w}$.

The cautious responses to a set of sender strategies S have to be computed separately (a) for the set of messages that occur in the range of some strategy in S, and (b) for the surprise messages, i.e. for those messages that are used by neither strategy in S.

For the first part, we construct for each receiver context c a utility matrix U^c for a three-person with the receiver, the sender, and the sender context as players. For a triple $(r \in \{r' \restriction_{(\bigcup_{s \in S_n} \text{range}(s))} \mid r' \in \mathcal{R}\}, s \in S, c \in \mathcal{C})$ (i.e. r is a function from the non-surprise messages to the actions), we have

$$U^c(r, s, c') = \sum_w p^*(w) u^R(c, w, r(c', s(c, w))).$$

All weakly undominated strategies in this game correspond to cautious responses to S (confined to the non-surprise messages), and vice versa.

Cautious responses to surprise messages can be computed separately for each surprise message. For each context c and surprise message f, we construct a two-person game $U^{c,f}$ with the actions as rows and the possible worlds in $[\![f]\!]$ as columns:

$$U^{c,f}(a, w) = u^R(c, w, a)$$

In each cautious response to S, (c, f) is mapped to some action that is weakly undominated in $U^{c,f}$.

The matrix games that figure in the computation of cautious responses to a set of sender strategies are usually quite large even for rather simple game. For instance, for a game with two worlds, two contexts, and two messages, we already have 16 sender strategies and 16 receiver strategies. So the three-person games that are needed to compute the set of cautious responses to some strategy set involves two matrices with up to $16 \times 16 \times 2 = 512$ cells each. It usually not practical to do these computations by hand. (In most cases, the matrices can be reduced considerably by first removing all strategies that are weakly dominated in pure strategies, but even this may be tedious to do by hand.)

However, there are efficient algorithms to find the weakly undominated strategies using Linear Programming (see for instance Conitzer and Sandholm 2005 for details). Therefore the ICR sequence can be computed in a rather straightforward way with the help of a computer.

This procedure will be spelled out in detail in connection with Example 6 below, which is both simple and revealing.

The considerable complexity of the ICR computation raises the question whether this is a realistic model of the reasoning processes that are performed in actual conversations. The ICR model belongs to the rationalistic tradition of game theory in this respect. It spells out how agents would behave in a given scenario if they were rational. It is well-known that real people are not rational in this sense. This insight does not necessarily invalidate the model though. There are three important considerations to be kept in mind here. First, rationality is an idealization just like the notions of frictionless motion or point masses in physics. Such idealizations do not represent reality in its entirety, but they capture relevant aspects of reality. As long as the interaction between real humans is governed by choices that can be approximated by the notion of rationality, a rationalistic game theoretic model provides a useful benchmark. Second, people learn from experience. Research on bounded rationality and on behavioral game theory has shown that under very general conditions, the outcome of a learning process may lead to a behavior that can be described as rational even if no rational reasoning is involved (see for instance Fudenberg and Levine 1998 on learning in games, and Camerer 2003 on experimental results). Third and finally, languages — or more specifically, the dispositions for linguistic behavior — constantly undergo a process of replication and selection that can be described by the logic of Darwinian evolution. As

research in evolutionary game theory has shown (see for instance Hofbauer and Sigmund 1998 for a fairly recent survey), such an evolutionary dynamics — just as many learning dynamics — frequently converge towards a disposition for apparently rational choices.

5 Examples

In the light of this formal definition, let us consider some of the previous examples again, which are repeated here for convenience.

Example 1 Completely aligned interests: We assume that for all signals f : $c(f) = 0$. There is only one context; v^S and u^R are given in Table 5.

	a_1	a_2
w_1	$1;1$	$0;0$
w_2	$0;0$	$1;1$

Table 5: Example 1

Here is the sequence of iterated computation of cautious responses, starting with the set R_0 of credulous strategies. The representation should be self-explanatory; every function that pairs each of the arguments in the left column with one of the arguments in the right column is part of the strategy set in question.

$$\mathbf{R} = R_0 = \begin{bmatrix} f_1 & \rightarrow & a_1 \\ f_2 & \rightarrow & a_2 \\ f_{12} & \rightarrow & a_1/a_2 \end{bmatrix}$$

$$\mathbf{S} = S_0 = \begin{bmatrix} w_1 & \rightarrow & f_1 \\ w_2 & \rightarrow & f_2 \end{bmatrix}$$

For a simple game such as this one, it is rather straightforward to compute the ICR sequence manually. The reasoning is as follows: In R_0, Robin's posterior probability distribution upon observing a message f is the uniform distribution over the set of worlds where f is true. For f_1 this is the singleton $\{w_1\}$. The action that maximizes his payoff in w_1 is a_1, hence $f_1 \rightarrow a_1$. The same reasoning applies to f_2. After observing f_{12}, both possible worlds are equally likely. Hence his expected payoff for both actions is $1/2$, and both are therefore best responses.

To compute S_0, we have to consider each possible world in turn. In w_0, Sally wants Robin to perform a_1. This can be achieved with probability 1 if she sends f_1 (and Robin uses R_0). Using f_{12} might also have this effect, but since both actions are here according to R_0, Sally will place some positive probability mass on both possible outcomes. So no matter which probability distribution over Robin's strategy she uses, the expected payoff for f_{12} will be lower than 1. Therefore f_1 is the only best response to R_0 in w_1. The same applies *ceteris paribus* to w_2.

	a_1	a_2
w_1	$1;-1$	$-1;1$
w_2	$-1;1$	$1;-1$

Table 6: Example 2

Example 2 Completely opposed interests: We still assume "cheap talk", i.e. all messages are costless. The utilities are repeated in Table 6

Here the iterative procedure enters a never-ending cycle:

$$
\begin{aligned}
R_0 &= \begin{bmatrix} f_1 & \rightarrow & a_2 \\ f_2 & \rightarrow & a_1 \\ f_{12} & \rightarrow & a_1/a_2 \end{bmatrix} \\
S_0 &= \begin{bmatrix} w_1 & \rightarrow & f_2 \\ w_2 & \rightarrow & f_1 \end{bmatrix} \\
R_1 &= \begin{bmatrix} f_1 & \rightarrow & a_1 \\ f_2 & \rightarrow & a_2 \\ f_{12} & \rightarrow & a_1/a_2 \end{bmatrix} \\
S_1 &= \begin{bmatrix} w_1 & \rightarrow & f_1 \\ w_2 & \rightarrow & f_2 \end{bmatrix} \\
R_2 &= R_0 \\
S_2 &= S_0 \\
&\vdots \\
\mathbf{R} &= \begin{bmatrix} f_1/f_2/f_{12} & \rightarrow & a_1/a_2 \end{bmatrix} \\
\mathbf{S} &= \begin{bmatrix} w_1/w_2 & \rightarrow & f_1/f_2 \end{bmatrix}
\end{aligned}
$$

So if the interests of the players are completely opposed, no communication will ensue.

Example 3 Rabin's example with partially aligned interests; the utilities are as in Table 7 and all signals are costless.

	a_1	a_2	a_3
w_1	$10;10$	$0;0$	$0;0$
w_2	$0;0$	$10;10$	$5;7$
w_3	$0;0$	$10;0$	$5;7$

Table 7: Example 3

$$
\begin{aligned}
R_0 &= \left[\begin{array}{lcl} f_1/f_{13} & \to & a_1 \\ f_2 & \to & a_2 \\ f_3/f_{23}/f_{123} & \to & a_3 \\ f_{12} & \to & a_1/a_2 \end{array}\right] \\
S_0 &= \left[\begin{array}{lcl} w_1 & \to & f_1/f_{13} \\ w_2/w_3 & \to & f_2 \end{array}\right] \\
R_1 &= \left[\begin{array}{lcl} f_1/f_{13} & \to & a_1 \\ f_2/f_3 & \to & a_3 \\ f_{12} & \to & a_1/a_2 \\ f_{23} & \to & a_2/a_3 \\ f_{123} & \to & a_1/a_2/a_3 \end{array}\right] \\
S_1 &= \left[\begin{array}{lcl} w_1 & \to & f_1/f_{13} \\ w_2/w_3 & \to & f_{12}/f_{23}/f_{123} \end{array}\right] \\
R_2 &= \left[\begin{array}{lcl} f_1/f_{13} & \to & a_1 \\ f_2 & \to & a_2 \\ f_3 & \to & a_3 \\ f_{12}/f_{23}/f_{123} & \to & a_2/a_3 \end{array}\right] \\
\neg(R_2(f_{12}) = R_2(f_{23}) &= R_2(f_{123}) = a_2)^{12}
\end{aligned}
$$

$$
\begin{aligned}
S_2 &= S_0 \\
R_3 &= R_1 \\
&\vdots \\
\mathbf{R} &= R_1 \cup R_2 \\
\mathbf{S} &= S_0 \cup S_1
\end{aligned}
$$

In S_0, Sally has the option to induce the desired outcome a_1 with certainty by using f_1 or f_{13}. Using f_{12} might also induce a_1, but it might also induce the sub-optimal a_2. As both options have a positive probability (recall that Sally's assumptions about Robin's behavior is an element of $\text{int}(\Delta(R_0))$, i.e. each element of R_0 has non-zero probability), choosing f_{12} in w_1 has a sub-optimal expected utility. The same kind of reasoning applies in w_2 and w_3, and analogously in the subsequent steps of the ICR sequence.

Note that no stable communication will emerge here in w_2 and w_3. Starting in S_0, Sally has the same set of options in w_2 and w_3. She may or may not choose to differentiate between w_2 and w_3; there are some cautious responses against R_1 that do and some that do not. Depending on Robin's private belief, he may expect to be able to differentiate between w_2 and w_3 on the basis of Sally's signal (and thus react to some signals with a_2), or he may prefer to play safe and choose a_3.

The situation changes drastically if the set of signals is confined to f_1 and f_{23}. Then we

have

$$\mathbf{R} = R_0 \;=\; \left[\begin{array}{lcl} f_1 & \to & a_1 \\ f_{23} & \to & a_3 \end{array}\right]$$

$$\mathbf{S} = S_0 \;=\; \left[\begin{array}{lcl} w_1 & \to & f_1 \\ w_2/w_3 & \to & f_{23} \end{array}\right].$$

Example 4 Next we will reconsider the example of the scalar implicature discussed above. Now we have two contexts, c_1 and c_2. The utilities are given in Table 8.

c_1 :	a_1	a_2	a_3
w_1	10; 10	0; 0	6; 6
w_2	0; 0	10; 10	6; 6

c_2 :	a_1	a_2	a_3
w_1	10; 10	0; 0	9; 9
w_2	0; 0	10; 10	9; 9

Table 8: Example 4

The signaling costs are as follows: $c(f_1) = c(f_{12}) = 0$ and $c(f_2) = 2$.[13]

$$R_0 \;=\; \left[\begin{array}{lcl} (c_1, f_1)/(c_2, f_1) & \to & a_1 \\ (c_1, f_2)/(c_2, f_2) & \to & a_2 \\ (c_1, f_{12})/(c_2, f_{12}) & \to & a_3 \end{array}\right]$$

$$S_0 \;=\; \left[\begin{array}{lcl} (c_1, w_1)/(c_2, w_1) & \to & f_1 \\ (c_1, w_2) & \to & f_2 \\ (c_2, w_2) & \to & f_{12} \end{array}\right]$$

$$\mathbf{R} = R_1 \;=\; \left[\begin{array}{lcl} (c_1, f_1)/(c_2, f_1) & \to & a_1 \\ (c_1, f_2)/(c_2, f_2)/(c_1, f_{12})/(c_2, f_{12}) & \to & a_2 \end{array}\right]$$

$$\mathbf{S} = S_1 \;=\; \left[\begin{array}{lcl} (c_1, w_1)/(c_2, w_1) & \to & f_1 \\ (c_1, w_2)/(c_2, w_2) & \to & f_{12} \end{array}\right]$$

The previous example illustrated how pragmatic rationalizability formalizes the intuition behind Levinson's (2000) **Q-Heuristics** "What isn't said, isn't." This heuristics accounts, *inter alia* for scalar implicatures such as the following:

(1) a. Some boys came in. $\rightsquigarrow$ Not all boys came in.

b. Three boys came in. $\rightsquigarrow$ Exactly three boys came in.

The essential pattern here is as in the schematic example above: There are two expressions A and B of comparable complexity such that the literal meaning of A entails the literal meaning of B. There is no simple expression for the concept "B but not A". In this scenario, a usage of "B" will implicate that A is false.

Example 5 Levinson assumes two further pragmatic principles that, together with the Q-principle, are supposed to replace Grice's maxims in the derivation of generalized conversational implicatures. The second heuristics, called **I-Heuristics**, says: "What is simply described is stereotypically exemplified." It accounts for phenomena of pragmatic strengthening, as illustrated in the following examples:

(3) a. John's book is good. $\rightsquigarrow$ The book that John is reading or that he has written is good.

b. a secretary $\rightsquigarrow$ a female secretary

c. road $\rightsquigarrow$ hard-surfaced road

The notion of "stereotypically exemplification" is somewhat vague and difficult to translate into the language of game theory. I will assume that propositions with a high prior probability are stereotypical. Also, I take it that "simple description" can be translated into "low signaling costs." So the principle amounts to "Likely propositions are expressed by cheap forms."

Let us construct a schematic example of such a scenario. Suppose there are two possible worlds (which may also stand for objects, like a hard surfaced vs. soft-surfaced road) w_1 and w_2, such that w_1 is *a priori* much more likely than w_2. Let us say that $p(w_1)/p(w_2) = 3$. There are three possible actions for Robin; he may choose a_1 if he expects w_1 to be correct, a_2 if he expects w_2, and a_3 if he finds it too risky to choose.

There are again three signals, f_1, f_2 and f_{12}. This time the more general expression f_{12} (corresponding for instance to "road") is cheap, while the two specific expressions f_1 and f_2 ("hard-surfaced road" and "soft-surfaced road") are more expensive: $c(f_1) = c(f_2) = 5$, and $c(f_{12}) = 0$.

The interests of Sally and Robin are completely aligned, except for the signaling costs which only matter for Sally. There are three contexts. In c_1 and c_2, it is safest for Robin to choose a_3 if he decides on the basis of the prior probability. In c_3 it makes sense to choose either a_1 or a_2 if he only knows the prior probabilities because the payoff of a_3 is rather low (but still higher than making the wrong choice between a_1 and a_2). In c_1, but not in c_2 it would be rational for Sally to use a costly message if this is the only way to make Robin perform a_1 rather than a_3. The precise utilities are given in Table 9.

$$R_0 = \begin{bmatrix} (c_1, f_1)/(c_2, f_1)/(c_3, f_1)/(c_3, f_{12}) & \rightarrow & a_1 \\ (c_1, f_2)/(c_2, f_2)/(c_3, f_2) & \rightarrow & a_2 \\ (c_1, f_{12})/(c_2, f_{12}) & \rightarrow & a_3 \end{bmatrix}$$

$$\mathbf{S} = S_0 = \begin{bmatrix} (c_1, w_1)/(c_3, w_1) & \rightarrow & f_1/f_{12} \\ (c_1, w_2)/(c_3, w_2) & \rightarrow & f_2 \\ (c_2, w_1) & \rightarrow & f_{12} \\ (c_2, w_2) & \rightarrow & f_2/f_{12} \end{bmatrix}$$

$$\mathbf{R} = R_1 = \begin{bmatrix} (c_1, f_1)/(c_2, f_1)/(c_3, f_1)/(c_3, f_{12}) & \rightarrow & a_1 \\ (c_1, f_2)/(c_2, f_2)/(c_3, f_2) & \rightarrow & a_2 \\ (c_1, f_{12})/(c_2, f_{12}) & \rightarrow & a_1/a_3 \end{bmatrix}$$

Here both f_1 and f_2 retain its literal meaning under pragmatic rationalizability. The unspecific f_{12} also retains its literal meaning in c_2. In c_1 and c_3, though, its meaning is pragmatically strengthened to $\{w_1\}$. Another way of putting is to say that f_{12} is *pragmatically*

c_1 :		a_1	a_2	a_3
	w_1	$28; 28$	$0; 0$	$22; 22$
	w_2	$0; 0$	$28; 28$	$22; 22$

c_2 :		a_1	a_2	a_3
	w_1	$28; 28$	$0; 0$	$25; 25$
	w_2	$0; 0$	$28; 28$	$25; 25$

c_3 :		a_1	a_2	a_3
	w_1	$28; 28$	$0; 0$	$10; 10$
	w_2	$0; 0$	$28; 28$	$10; 10$

Table 9: Example 5

ambiguous here. Even though it has an unambiguous semantic meaning, its pragmatic interpretation varies between contexts. It is noteworthy here that f_{12} can never be strengthened to mean $\{w_2\}$. Applying it to the example, this means that a simple non-specific expression such as "road" can either retain its unspecific meaning, or it can be pragmatically strengthened to its stereotypical instantiation (such as *hard-surfaced road* here). It can never be strengthened to a non-stereotypical meaning though.

Example 6 Levinson's third heuristics is the **M-heuristics**: "What is said in an abnormal way isn't normal." It is also known, after Horn (1984), as **division of pragmatic labor**. A typical example is the following:

(4) a. John stopped the car.

b. John made the car stop.

The two sentences are arguably semantically synonymous. Nevertheless they carry different pragmatic meanings if uttered in a neutral context. (4a) is preferably interpreted as *John stopped the car in a regular way, like using the foot brake.* This would be another example for the I-heuristics. (4b), however, is also pragmatically strengthened. It means something like *John stopped the car in an abnormal way, like driving it against a wall, making a sharp u-turn, driving up a steep mountain, etc.*

This can be modeled quite straightforwardly. Suppose there are again two worlds, w_1 and w_2, such that w_1 is likely and w_2 is unlikely (such as using the foot brake versus driving against a wall). Let us say that $p(w_1)/p(w_2) = 3$ again. There are two actions, a_1 and a_2, which are best responses in w_1 and w_2 respectively. There is only one context. The utilities are given in Table 10.

Unlike in the previous example, we assume that there are only two expressions, f and f', which are both unspecific: $[\![f]\!] = [\![f']\!] = \{w_1, w_2\}$. (Or, alternatively, we might assume that f_1 and f_2 are prohibitively expensive.) f' is slightly more expensive than f, like $c(f) = 0$ and $c(f') = 1$.

	a_1	a_2
w_1	$5;5$	$0;0$
w_2	$0;0$	$5;5$

Table 10: Example 6

$$
\begin{aligned}
R_0 &= \left[\ f/f' \ \rightarrow \ a_1\ \right] \\
S_0 &= \left[\ w_1/w_2 \ \rightarrow \ f\ \right] \\
R_1 &= \left[\begin{array}{lll} f & \rightarrow & a_1 \\ f' & \rightarrow & a_1/a_2 \end{array}\right] \\
S_1 &= \left[\begin{array}{lll} w_1 & \rightarrow & f \\ w_2 & \rightarrow & f/f' \end{array}\right] \\
\mathbf{R} = R_2 &= \left[\begin{array}{lll} f & \rightarrow & a_1 \\ f' & \rightarrow & a_2 \end{array}\right] \\
\mathbf{S} = S_2 &= \left[\begin{array}{lll} w_1 & \rightarrow & f \\ w_2 & \rightarrow & f' \end{array}\right]
\end{aligned}
$$

The crucial point here is that in S_0, the signal f' remains unused. Therefore any rationalizable interpretation of f' which is compatible with its literal meaning is licit in R_1, including the one where f' is associated with w_2 (which triggers the reaction a_2). Robin's reasoning at this stage can be paraphrased as: If Sally uses f, this could mean either w_1 or w_2. Since w_1 is *a priori* more likely, I will choose a_1. There is apparently no good reason for Sally to use f'. If she uses it nevertheless, she must have something in mind which I hadn't thought of.[14] Perhaps she wants to convey that she is actually in w_2.

Sally in turn reasons: If I say f, Robin will take action a_1. If I use f', he may take either action. In w_1 I will thus use f. In w_2 I can play it safe and use f, but I can also take my chances and try f'.

Robin in turn will calculate in R_2: If I hear f, we are in w_1 with a confidence between 75% and 100%. In any event, I should use a_1. The only world where Sally would even consider using f' is w_2. So if I hear f', the posterior probability of w_2 is 100%, and I can safely choose a_2.

If Robin reasons this way, it is absolutely safe for Sally to use f' in w_2.

Digression: Algorithmic computation of the ICR sequence Let me use this example to illustrate the algorithmic procedure to compute the ICR sequence that was presented at the end of Section 4.

The expected utility for Robin for choosing action a as response to either signal in R_0 is $\sum_{w \in \mathcal{W}} u^R(w, a)$, as both f and f' are true in all worlds. For a_1 this is $15/4$, and for a_2 $5/4$. Therefore R_0 will choose a_1 for each action.

To compute S_0, we construct a matrix U^w for each world w. (As there is only one context in this game, c can be omitted.) As R_0 is a singleton set, this is trivial. Let r be the only member of R_0:

$$U^{w_1} = \begin{array}{c|c} & r \\ \hline f & 5 \\ f' & 4 \end{array} \quad U^{w_2} = \begin{array}{c|c} & r \\ \hline f & 0 \\ f' & -1 \end{array}$$

As f dominates f' in both matrices, S_0 always uses f.

To compute R_1, first the interpretation of the only non-surprise message, f, is considered. The rows are the possible sender strategies, confined to the non-surprise messages. As there is only one such message here, this amounts to the set of actions. The columns are the sender strategies in S_0. There is only one of those, calls it s.

$$U^f = \begin{array}{c|c} & s \\ \hline a_1 & 15/4 \\ a_2 & 5/4 \end{array}$$

a_1 dominates a_2. Therefore f is mapped to a_1 in S_1.

As next step, we construct a matrix for each surprise message, with actions as rows and the worlds where this message is true as columns. This only applies to f' here:

$$U^{f'} = \begin{array}{c|cc} & w_1 & w_2 \\ \hline a_1 & 5 & 0 \\ a_2 & 0 & 5 \end{array}$$

Neither row dominates the other one, so both actions are cautious responses to S_0.

Now we have two receiver strategies in R_1: r_{11} (mapping both f and f' to a_1) and r_{12} (mapping f to a_1 and f' to a_2). So the two matrices for the computations of S_1 are:

$$U^{w_1} = \begin{array}{c|cc} & r_{11} & r_{12} \\ \hline f & 5 & 5 \\ f' & 4 & -1 \end{array} \quad U^{w_2} = \begin{array}{c|cc} & r_{11} & r_{12} \\ \hline f & 0 & 0 \\ f' & -1 & 4 \end{array}$$

As a response to w_1, f dominates f'. In w_2, both messages are undominated. So we have two sender strategies in S_1: s_{11} (mapping both worlds to f) and s_{12} (mapping w_1 to f and w_2 to f'). As there are no surprise messages in S_1, the matrix for computing R_2 is:

$$U = \begin{array}{c|cc} & s_{11} & s_{12} \\ \hline r_{11} & 15/4 & 15/4 \\ r_{12} & 15/4 & 5 \\ r_{21} & 5/4 & 0 \\ r_{22} & 5/4 & 5/4 \end{array}$$

Row r_{12} strictly dominates r_{21} and r_{22}, and it weakly dominates r_{11}. Therefore r_{12} is the only cautious response to S_1.

The computation of S_2 is rather trivial because there is only one receiver strategy in R_2:

$$U^{w_1} = \begin{array}{c|c} & r_{12} \\ \hline f & 5 \\ f' & -1 \end{array} \quad U^{w_2} = \begin{array}{c|c} & r_{12} \\ \hline f & 0 \\ f' & 4 \end{array}$$

Clearly f dominates f' in w_1, and f' dominates f in w_2.

Example 7 M-implicatures have been used as motivating example for **bidirectional Optimality Theory** (see for instance Blutner 2001) as a framework for formal pragmatics. It has been shown in Jäger (2002) that the set of (weakly) bidirectionally optimal form-meaning pairs can be computed by an iterative procedure that has some similarity to the one given in Definition 4. It is thus an interesting questions how the two frameworks relate.[15]

Weak bidirectionality predicts that simple forms are paired with stereotypical meanings and complex forms with atypical meanings. The prediction is even stronger though: if the set of forms in question can be ordered according to complexity in a linear way, such as $c(f_1) < c(f_2) < \cdots < c(f_n)$, and the set of meanings has the same cardinality and can also be ordered in a linear fashion (such as $p(w_1) > p(w_2) > \cdots > p(w_n)$, then the bidirectionally optimal pairs are all pairs (f_i, w_i).

Let us see what pragmatic rationalizability predicts. Suppose there are three worlds with $p(w_1) > p(w_2) > p(w_3)$. Also, there are three forms with $c(f) < c(f') < c(f')$ which are semantically synonymous, namely $[\![f]\!] = [\![f']\!] = [\![f'']\!] = \{w_1, w_2, w_3\}$. There are three actions such that exactly one action is optimal for each world for both players. There is only one context; the utilities are as in Table 11.

	a_1	a_2	a_3
w_1	5; 5	0; 0	0; 0
w_2	0; 0	5; 5	0; 0
w_3	0; 0	0; 0	5; 5

Table 11: Example 7

Here is the iterative reasoning sequence:

$$R_0 = \left[\begin{array}{lcl} f/f'/f'' & \to & a_1 \end{array} \right]$$

$$S_0 = \left[\begin{array}{lcl} w_1/w_2/w_3 & \to & f \end{array} \right]$$

$$R_1 = \left[\begin{array}{lcl} f & \to & a_1 \\ f'/f'' & \to & a_1/a_2/a_3 \end{array} \right]$$

$$\mathbf{S} = S_1 = \left[\begin{array}{lcl} w_1 & \to & f \\ w_2/w_3 & \to & f/f'/f'' \end{array} \right]$$

$$\mathbf{R} = R_2 = \left[\begin{array}{lcl} f & \to & a_1 \\ f'/f'' & \to & a_2/a_3 \end{array} \right]$$

Pragmatic rationalizability makes significantly weaker predictions than bidirectional OT. We do predict a division of pragmatic labor in the sense that the cheapest form, f, is specialized to the most probable interpretation w_1 (and the corresponding best action a_1), while the more complex forms f' and f'' are specialized to the non-stereotypical meanings. However, no further specialization between f' and f'' is predicted.

This seems to be in line with the facts. Next to the two expressions in (4), there is a third alternative, which is still more complex than (4b).

(5) John brought the car to a stop.

Also, there are various non-standard ways of making a car stop. The most probable way besides using the foot brake is perhaps to use the hand brake, driving against a wall is even less likely. So bidirectional OT would predict that (4b) carries the implicature that John used the hand brake, while (5) is restricted to even more unusual ways of stopping a car. The present framework only predicts that both (4b) and (5) convey the information that John acted in a somehow non-stereotypical way. While intuitions are not very firm here, it seems to me that the predictions of bidirectional OT might in fact be too strong here.

Example 8 Here is another example that has been analyzed by means of bidirectional OT in the literature. Krifka (2002) observes that the pragmatic interpretation of number words follows an interesting pattern that is reminiscent of Levinson's M-heuristics:

> "RN/RI principle:
>
> a. Short, simple numbers suggest low precision levels.
>
> b. Long, complex numbers suggest high precision levels."
>
> (Krifka 2002:433)

This can be illustrated with the following contrast:

(6) a. The distance is one hundred meter.

b. The distance is one hundred and one meter.

The sentence (6b) suggests a rather precise interpretation (with a slack of at most 50 cm), while (6a) can be more vague. It may perhaps mean something between 90 and 110 meter. Actually, (6a) is pragmatically ambiguous; depending on context, it can be rather precise or rather vague. The crucial observation here is: A shorter number term such as "one hundred" allows for a larger degree of vagueness than a more complex term such as "one hundred and one."

Krifka also observes that the degree of vagueness of a short term can be reduced by making it more complex — for instance by modifying it with "exactly":

(7) The distance is exactly one hundred meter.

Krifka (2002) accounts for these facts in terms of bidirectional OT, assuming a general preference for vague over precise interpretation. Krifka (2007) contains a revised analysis which employs game theoretic pragmatics. Space does not permit a detailed discussion of

Krifka's proposals; in the following I will just briefly sketch how pragmatic rationalizability accounts for Krifka's observations.

Suppose there are two equiprobable worlds, w_1 and w_2. Suppose the distance in question is exactly 100 meter in w_1 and 101 meter in w_2. There are three signals: f_1 ("The distance is one hundred meter."), f_1' ("The distance is exactly one hundred meter.") and f_2 ("The distance is one hundred and one meter."). So we have $[\![f_1]\!] = [\![f_1']\!] = \{w_1\}$, and $[\![f_2]\!] = \{w_2\}$. Let us assume that $c(f_1) = 0$ and $c(f_1') = c(f_2) = 4.5$. There are two actions. a_1 is optimal for w_1 and a_2 for w_2. Furthermore, there are two contexts. In c_1, precision is very important. This means that the differential costs of using an expensive message are lower than the difference in utility between a_1 and a_2. In c_2 it is the other way round.

Table 12 gives the numerical utilities:

		a_1	a_2
c_1 :	w_1	$10; 10$	$0; 0$
	w_2	$0; 0$	$10; 10$

		a_1	a_2
c_2 :	w_1	$4; 4$	$0; 0$
	w_2	$0; 0$	$4; 4$

Table 12: Example 8

Here is the iterative reasoning sequence:

$$R_0 = \begin{bmatrix} (c_1, f_1) & \rightarrow & a_1 \\ (c_1, f_1') & \rightarrow & a_1 \\ (c_1, f_2) & \rightarrow & a_2 \\ (c_2, f_1) & \rightarrow & a_1 \\ (c_2, f_1') & \rightarrow & a_1 \\ (c_2, f_2) & \rightarrow & a_2 \end{bmatrix}$$

$$S_0 = \begin{bmatrix} (c_1, w_1)/(c_2, w_1)/(c_2, w_2) & \rightarrow & f_1 \\ (c_1, w_2) & \rightarrow & f_2 \end{bmatrix}$$

$$\mathbf{R} = R_1 = \begin{bmatrix} (c_1, f_1) & \rightarrow & a_1/a_2 \\ (c_1, f_1') & \rightarrow & a_1 \\ (c_1, f_2) & \rightarrow & a_2 \\ (c_2, f_1) & \rightarrow & a_1/a_2 \\ (c_2, f_1') & \rightarrow & a_1 \\ (c_2, f_2) & \rightarrow & a_2 \end{bmatrix}$$

$$\mathbf{S} = S_1 = \begin{bmatrix} (c_1, w_1) & \rightarrow & f_1/f_1' \\ (c_1, w_2) & \rightarrow & f_2 \\ (c_2, w_1)/(c_2, w_2) & \rightarrow & f_1 \end{bmatrix}$$

The two complex expressions f_2 and f_1' are always interpreted in a precise way under the PRSs. The simple expression f_1 is pragmatically ambiguous between a precise interpretation (in c_1) and a vague interpretation (in c_2).

6 Comparison to Franke's IBR model: scalar implicatures again

In a series of recent publications, Michael Franke has developed the **Iterated Best Response model** (IBR model) of game theoretic pragmatics (see Franke 2009, 2011). This model is conceptually very similar to the present model, so let me briefly discuss where the models differ.

Franke gives a detailed procedure how a linguistic example is to be transformed into a game. In these games, the actions are always isomorphic to the possible worlds (or *types*, as he calls it), and the utility function is based on *type matching*, i.e. both players score 1 if Robin picks the action that corresponds to the correct world, and 0 otherwise. He also considers so-called *epistemic games* where types are sets of possible worlds, i.e. information states. This models scenarios where Sally is not fully informed (or, to be precise, where it is not common knowledge that she is).

These design decisions are of course fully compatible with the ICR model (see Jäger 2011 for a variant of the ICR model where incomplete knowledge of the sender about the true possible world is incorporated). The crucial difference lies in the way a response to a non-singleton set of possible opposing strategies are computed. While the ICR model considers all possible probability distributions over the set of possible strategies (as long as they assign positive probability to all possibilities), Franke assumes that the *Principle of Insufficient Reason* (see Jaynes 2003) applies. This means that all possibilities receive the same probability. A best response to a set of strategies is thus conceived as a best response to the uniform distribution over these strategies.

This corresponds to a subtle difference in the epistemic foundations of the two models. In Franke's model, the agents have exactly the same amount of information as the modeler. If, at a certain stage of the iterative reasoning process, the opposing player has more than one option, the reasoning agent has no reason to prefer any of these options over another. In the present model, it is very well possible that the agents have certain prior assumptions about the dispositions of the other player. The ICR model computes the predictions that we can make if we do not know these prior assumptions.

This can be illustrated with an abstract example. Reconsider the utility matrix from Example 6, which is repeated here for convenience: As in Example 6, we assume that there

	a_1	a_2
w_1	$5;5$	$0;0$
w_2	$0;0$	$5;5$

Table 13: Example 9

are two messages, f and f', which are both true in both worlds. However, we now assume that both worlds are equally likely and that both messages are costless.

The ICR sequence comes out as:

$$\mathbf{R} = R_0 \;=\; \left[\; f/f' \;\rightarrow\; a_1/a_2 \;\right]$$

$$\mathbf{S} = S_0 \;=\; \left[\; w_1/w_2 \;\rightarrow\; f/f' \;\right]$$

In Franke's model, (R_0, S_0) is likewise a fixed point. His R_0 is a mixed strategy though where Robin performs both actions with equal probability, no matter which signal is observed. Likewise, his S_0 is a mixed strategy where Sally sends either message with 50% probability, regardless of the world she is in.

While both models agree that it is not possible to predict any reliable information exchange in this scenario, the interpretation of this result is different. Franke's model predicts that in fact no information transmission takes place. It is fully consistent with the ICR result though that Sally uses f in w_1 and f' in w_2 with a probability close to 1, and that Robin interprets f as a_1 and f' as a_2, also with a probability close to 1. It is thus consistent with the predictions of the ICR model that a differentiation of meanings takes place — it is just not possible to predict from the available information how it looks like.

While this consideration may be rather abstract and meta-theoretical, the following example points to a more substantial difference.

Franke (2009) (page 77 pp.) construes the simple scalar implicature scenario as the following game. We have two possible worlds, $w_\forall$ (where all boys came to the party) and $w_{\exists\neg\forall}$ (where some but not all boys came to the party). There are three messages, f_s ("Some boys came to the party.") with $[\![f_s]\!] = \{w_\forall, w_{\exists\neg\forall}\}$, f_a ("All boys came to the party.") with $[\![f_a]\!] = \{w_\forall\}$, and f_{sbna} ("Some but not all boys came to the party") with $[\![f_{sbna}]\!] = \{w_{\exists\neg\forall}\}$. The latter signal incurs a small cost, while the other two messages are costless. The utility matrix is based on type matching. It is shown in Table 14.

	$a_\forall$	$a_{\exists\neg\forall}$
$w_\forall$	1; 1	0; 0
$w_{\exists\neg\forall}$	0; 0	1; 1

Table 14: Example 10

Here the ICR sequence comes out as shown in the left panel of Table 15. Consider the computation of S_0. In $w_\forall$, the optimal outcome of $a_\forall$ can be induced with certainty by using f_a, and with an unknown probability in $(0, 1)$ by using f_s. As both signals are equally cheap, the former is the safer bet. In $w_{\exists\neg\forall}$, however, f_s might be the better option because it is cheaper than f_{sbna}. If Sally considers it sufficiently likely that Robin will map f_s to $a_{\exists\neg\forall}$, it is rational for her to use this message, because the risk of being misunderstood is offset by the reduced message costs. If Sally assumes that f_s will be mapped to $a_\forall$ with sufficiently high probability, it is better for her to use f_{sbna}.

Due to this indeterminacy, f_{sbna} is not a surprise message in S_0. Therefore Robin will conclude with certainty that Sally is in $w_{\exists\neg\forall}$ when observing it, and accordingly choose $a_{\exists\neg\forall}$.

In Franke's model, f_s will be mapped to both actions with equal likelihood in R_0. Sally's expected payoff for using it in $w_{\exists\neg\forall}$ is thus 0.5. Using f_{sbna} will induce $a_{\exists\neg\forall}$ with probability 1. If the costs of f_{snba} are smaller than 0.5, she will therefore choose it. f_s comes out as a surprise message in S_0.

In the ICR model, it is assumed that the only information that Robin draws from observing a surprise message is that this message is true. If we apply this principle in connection with the Principle of Insufficient reason, we get the IBR sequence as shown in the left panel of Table 15. Both $a_\forall$ and $a_{\exists\neg\forall}$ are equally good responses to f_s under the assumption that f_s is true. Therefore the best response to S_0 is again R_0.

$$R_0 = \begin{bmatrix} f_s & \to & a_\forall / a_{\exists\neg\forall} \\ f_a & \to & a_\forall \\ f_{sbna} & \to & a_{\exists\neg\forall} \end{bmatrix} \qquad \mathbf{R} = R_0 = \begin{bmatrix} f_s & \to & a_\forall / a_{\exists\neg\forall} \\ f_a & \to & a_\forall \\ f_{sbna} & \to & a_{\exists\neg\forall} \end{bmatrix}$$

$$S_0 = \begin{bmatrix} w_\forall & \to & f_a \\ w_{\exists\neg\forall} & \to & f_s / f_{sbna} \end{bmatrix} \qquad \mathbf{S} = S_0 = \begin{bmatrix} w_\forall & \to & f_a \\ w_{\exists\neg\forall} & \to & f_{sbna} \end{bmatrix}$$

$$\mathbf{R} = R_1 = \begin{bmatrix} f_s & \to & a_{\exists\neg\forall} \\ f_a & \to & a_\forall \\ f_{sbna} & \to & a_{\exists\neg\forall} \end{bmatrix}$$

$$\mathbf{S} = S_0 = \begin{bmatrix} w_\forall & \to & f_a \\ w_{\exists\neg\forall} & \to & f_s \end{bmatrix}$$

Table 15: Comparison ICR vs. IBR

Franke discusses this problem at length. His solution is to adopt a more sophisticated method for interpreting surprise messages. Briefly put, when Robin observes a surprise messages, he wonders in which world Sally could possibly benefit from deviating from her strategy. In the current example, Sally will achieve her maximal payoff anyway when she is in $w_\forall$. In $w_{\exists\neg\forall}$, she might hope to save message costs by sending a costless surprise message. Therefore Robin's best explanation for observing f_s is that Sally is in $w_{\exists\neg\forall}$, and he will accordingly choose $a_{\exists\neg\forall}$.

To sum up, Franke's model uses a coarser method for choosing a response when the strategy of the opposing player is not known. This reduces the complexity of the computations considerably. On the other hand, it may lead to unwelcome results as soon as message costs enter the picture. To remedy this problem, Franke has to adopt a more complex belief revision policy for the interpretation of surprise messages.

It should be noted that Franke's model (including this belief revision policy) makes the same predictions as bidirectional OT in scenarios such as Example 7.

7 Related work

The essential intuition behind the proposal laid out here is that the literal meaning of signals constitutes their default interpretation, and that rational communicators decide about their communicative strategies by iteratively calculating the best response to this default strategy. Similar ideas have been proposed at various places in the literature, sometimes implicitly, even though the precise technical implementation offered here is to my knowledge novel.

As briefly discussed above, Rabin (1990) gives a definition when a message should count as credible. Within the present framework, his definition could be recast as: a message f is credible iff for each n and for each $s \in S_n$, $[\![f]\!] \subseteq s^{-1}(f)$. This equivalence only holds under certain side conditions pertaining to the space of available messages, but essentially Rabin's definition of credibility relies on an iterated calculation of best responses, starting with the credulous receiver strategies.

Stalnaker (2005) proposes an informal notion of credibility that could be interpreted as

follows: f is credible iff there is some $s \in S_0$ such that $f \in \text{range}(s)$, and for each $s \in S_0$: $s^{-1}(f) \subseteq [\![f]\!]$.

Benz and van Rooij (2007) develop a pragmatic interpretation procedure that can, in the present framework, be approximated by the rule: Sally should choose her signals according to S_0, and Robin should interpret them according to R_1. They assume an additional constraint though requiring that only honest strategies will be admitted in S_0.

In Jäger (2007) I propose to calculate the pragmatically licit communication strategies by starting with a strategy based on the literal interpretation of signals and iteratively computing the best response strategy until a fixed point is reached. So this approach is similar in spirit to the present one to a certain degree. Nevertheless the two theories differ considerably in detail. In Jäger (2007) I assumed an update rule where S_{n+1}/R_{n+1} are mixed strategies that differ only infinitesimally from S_n/R_n. The reasoning process that is modeled this way is quite unlike the Gricean inference schemes that are dealt with in the present framework.

The present paper is a revised version of a manuscript that was written in 2008 and has been in circulation under the title "Game Theory in Semantics and Pragmatics" since then. Some of the ideas laid out there have been taken up and developed further in subsequent work such as Franke (2009), Jäger and Ebert (2009), Franke (2011), Jäger (2011), and Franke and Jäger (2012). The most significant innovation here arguably is Franke's usage of the Principle of Insufficient Reason to narrow down the space of strategies at each level of the iterative reasoning sequence to a single strategy.

8 Conclusion

This article is primarily aimed at introducing readers with a background in linguistic semantics and pragmatics to some of the issues that game theorists worry about when study the conditions for communication between rational agents. At the same time, the article might be of use for economists and philosophers with a background in game theory that are interested in the specific problems of linguistic pragmatics and the potential of game theoretic methods in this domain.

The question whether or not it is rational to communicate at all in a particular situation has largely been ignored in the linguistic research tradition because a complete alignment of interests is usually assumed. The game theoretic research has shown that communication can be rational even if the interests of the interlocutors are only partially aligned.

A second issue that is prominent in the game theoretic discussion is the role of conventionalized meaning of messages in situations where a simple-minded assumption of honesty and credulity is in partial conflict with rationality. This is also one of the core concerns of Gricean pragmatics. I proposed a game theoretic formalization of Gricean reasoning that both captures the intuitive reasoning patterns that are traditionally assumed in the computation of implicatures, and that addresses the problem of the credibility of signals under partially aligned interests of the interlocutors.

9 Acknowledgments

Thanks to Michael Franke and Robert van Rooij for a very illuminating discussion, to Michael Franke and Christian Ebert for pointing out several mistakes in a draft version of this article, and to the editor of this special issue as well as two anonymous reviewers for helpful comments.

Notes

[1]The analogous problem also arises in van Rooij's model.

[2]In the game-theoretic terminology, both L_1 and L_2 constitute strict Nash equilibria.

[3]A strategy s is *rationalizable* if there is a consistent set of beliefs such that s maximizes the expected payoff of the player, given these beliefs and the assumption that rationality of all players is common knowledge.

[4]The decision to use integers for sender types and real numbers for actions is purely out of mathematical convenience.

[5]The first step follows from a standard result in statistics, according to which the best estimator for a squared-error loss function — such as u^R in the example — is the expected value (see for instance Jaynes 2003, p. 416). The second step utilizes the facts that $\sum_{w=1}^{\infty} 2^{-w} = 1$ and $\sum_{w=1}^{\infty} w \cdot 2^{-w} = 2$, which can both easily be shown via complete induction.

[6]Thanks to Michael Franke for pointing this out to me.

[7]In many scenarios, the intuitions about what constitutes a credible message is somewhat less clear than in the ones presented here. This has led to a lively debate about how credibility should be precisely defined. The interested reader is referred to Rabin (1992); Farrell (1993); Farrell and Rabin (1996); Zapater (1997); Stalnaker (2005) and the literature cited therein.

[8]There might be more than one credulous strategy because several actions may yield the same maximal payoff for Robin in certain situations.

[9]Grice's Maxim of Quantity is " Make your contribution as informative as is required (for the current purposes of the exchange). Do not make your contribution more informative than is required.", and the Maxim of Manner "Be Clear. Avoid obscurity of expression. Avoid ambiguity. Be brief (avoid unnecessary prolixity). Be orderly." (cf. Grice 1975).

[10]I use the term "context" in such a way here that the preferences of the players may vary between contexts (as well as between worlds), while the literal meaning of messages is invariant between contexts. So this notion of context has nothing to do with the knowledge state of the discourse participants or the interpretation of indexical expressions.

[11]Epistemically speaking, this means that I do not assume any common belief about which context the players are in, even though they might hold private beliefs.

[12]Note that in $S2$, both in w_2 and w_3 Sally sends one of the messages f_{12}, f_{23} and f_{123}. Robin assumes that in each world, Sally uses some probability distribution over these messages. If he thinks that there is one message that is sufficiently more likely to be sent in w_2 than in w_3, this message will induce a high posterior probability for w_2, which in turn makes it rational for him to pick a_2. It is not possible though that all three messages are more likely to be sent in w_2 than in w_3. Therefore it is never rational for Robin to pick a_2 for all three messages.

[13]One reviewer remarked that the usage of significant signaling costs is somewhat unusual as compared to most of the signaling game literature. The standard assumption is that costs are *nominal*, i.e. that they are vanishingly small in comparison to other determinants of utilities, and that they only play a role as tie-breakers. I find the assumption of nominal signaling costs unrealistic though in the context of linguistic communication. There is always a certain limit to the tolerable complexity of the expression that is being used, so at some point speakers will always sacrifice precision for brevity.

[14]The idea that Robin is prepared to revise his prior assumptions in any arbitrary way upon observing a surprise message is technically implemented by *existentially quantifying* over Robin's probability distribution p in the last line of Definition 4, rather than using his prior distribution p^*.

[15]See Franke and Jäger (2012) for a detailed discussion of the relation between bidirectional OT and game theoretic reasoning.

References

Benz, Anton and Robert van Rooij. 2007. Optimal assertions and what they implicate. *Topoi — an International Review of Philosophy* 27, 63–78.

Blutner, Reinhard. 2001. Some aspects of optimality in natural language interpretation. *Journal of Semantics* 17, 189–216.

Camerer, Colin F. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction.* Princeton: Princeton University Press.

Conitzer, Vincent and Thomas Sandholm. 2005. Complexity of (iterated) dominance. *Proceedings of the 6th ACM conference on Electronic commerce*, 88–97, ACM.

Crawford, Vincent P. and Joel Sobel. 1982. Strategic Information Transmission. *Econometrica* 50, 1431–1451.

Farrell, Joseph. 1993. Meaning and credibility in cheap-talk games. *Games and Economic Behavior* 5, 514–531.

Farrell, Joseph and Matthew Rabin. 1996. Cheap talk. *The Journal of Economic Perspectives* 10, 103–118.

Franke, Michael. 2009. Signal to act: Game theory in pragmatics. Ph.D. thesis, University of Amsterdam.

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

Franke, Michael and Gerhard Jäger. 2012. Bidirectional optimization from reasoning and learning in games. *Journal of Logic, Language and Information* 21, 117–139.

Fudenberg, Drew and David Levine. 1998. *The Theory of Learning in Games*. Cambridge, Mass.: MIT Press.

Grice, Herbert Paul. 1975. Logic and conversation. *Syntax and Semantics 3: Speech Acts*, edited by P. Cole and J. Morgan, 41–58, New York: Academic Press.

Hofbauer, Josef and Karl Sigmund. 1998. *Evolutionary Games and Population Dynamics*. Cambridge, UK: Cambridge University Press.

Horn, Laurence. 1984. Towards a new taxonomy for pragmatic inference: Q-based and R-based implicatures. *Meaning, Form, and Use in Context*, edited by Deborah Schiffrin, 11–42, Washington: Georgetown University Press.

Jäger, Gerhard. 2002. Some notes on the formal properties of bidirectional Optimality Theory. *Journal of Logic, Language and Information* 11, 427–451.

Jäger, Gerhard. 2007. Game dynamics connects semantics and pragmatics. *Game Theory and Linguistic Meaning*, edited by Ahti-Veikko Pietarinen, 89–102, Elsevier.

Jäger, Gerhard. 2011. Game-theoretical pragmatics. *Handbook of Logic and Language*, edited by Johan van Benthem and Alice ter Meulen, 467–491, Elsevier, 2nd edition.

Jäger, Gerhard and Christian Ebert. 2009. Pragmatic rationalizability. *Proceedings of Sinn und Bedeutung 13*, edited by Arndt Riester and Torgrim Solstad, number 5 in SinSpeC. Working Papers of the SFB 732, 1–15, University of Stuttgart.

Jaynes, Edwin Thompson. 2003. *Probability Theory: The Logic of Science*. Cambridge University Press.

Krifka, Manfred. 2002. Be brief and vague! and how bidirectional optimality theory allows for verbosity and precision. *Sounds and Systems. Studies in Structure and Change. A Festschrift for Theo Vennemann*, edited by David Restle and Dietmar Zaefferer, 439–358, Berlin: Mouton de Gruyter.

Krifka, Manfred. 2007. Approximate interpretation of number words: A case for strategic communication. *Cognitive foundations of interpretation*, edited by Gerlof Bouma, Irene Krämer, and Joost Zwarts, 111–126, Amsterdam: Koninklijke Nederlandse Akademie van Wetenschapen.

Levinson, Stephen C. 2000. *Presumptive Meanings*. MIT Press.

Lewis, David. 1969. *Convention*. Cambridge, Mass.: Harvard University Press.

Osborne, Martin J. 2003. *An Introduction to Game Theory*. Oxford: Oxford University Press.

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

Parikh, Prashant. 2010. *Language and Equilibrium*. Cambridge (Mass.): MIT Press.

Pearce, David G. 1984. Rationalizable strategic behavior and the problem of perfection. *Econometrica* 52.

Rabin, Matthew. 1990. Communication between rational agents. *Journal of Economic Theory* 51, 144–170.

Rabin, Matthew. 1992. Corrigendum. *Journal of Economic Theory* 58, 110–111.

Stalnaker, Robert. 2005. Saying and meaning, cheap talk and credibility. *Game Theory and Pragmatics*, edited by Anton Benz, Gerhard Jäger, and Robert van Rooij, 83–100, Palgrave MacMillan.

van Rooij, Robert. 2004. Signalling games select Horn strategies. *Linguistics and Philosophy* 27, 493–527.

Zapater, Iñigo. 1997. Credible proposals in communication games. *Journal of Economic Theory* 72, 173–197.