Optimal Assertions, and what they Implicate.

A uniform game theoretic approach.

Anton Benz & Robert van Rooij*

Abstract

To determine what the speaker in a cooperative dialog meant with his assertion, on top of what he explicitly said, it is crucial that we assume that the assertion he gave was optimal. In determining optimal assertions we assume that dialogues are embedded in decision problems (van Rooij, 2003) and use backwards induction for calculating them (Benz, 2006). In this paper we show that in terms of our framework we can account for several types of implicatures in a uniform way, suggesting that there is no need for an independent linguistic theory of generalized implicatures. In the final section we show how we can embed our theory in the framework of signaling games, and how it relates with other game theoretic analyses of implicatures.

1 Introduction

The study of language is traditionally divided into three domains: syntax, semantics, and pragmatics. *Syntax* is the study of expressions independent of what they mean, while *semantics* studies the conventional meaning of these expressions. This paper belongs to the realm of *pragmatics*, which studies the interpretation of expressions in the particular context in which they are used. Perhaps the most important notion in linguistic pragmatics is Grice's (1967) notion of *conversational implicature*. It is based on the insight that by means of general principles of rational cooperative communication we can communicate more with the *use* of a sentence than the *conventional* semantic *meaning* associated with it. What is communicated by the use of an expression or sentence depends not only on syntactic and semantic rules, but also on some basic assumptions about the rational nature of conversational activity and the preferences and expectations of the agents involved. In general, the interpretation of an utterance depends on what the speaker expects the hearer will understand, which in turn depends, in a circular way, on what the listener thinks that the speaker has in mind. Thus, communication requires coordination between speaker and hearer which involves interactive reasoning. As *game theory* is the general study of interactive reasoning, it is only natural to assume – following Lewis' (1969) game theoretic analysis

*The names of the authors are listed in alphabetic order. We would like to thank Bart Geurts, Wilfrid Hodges, Kris de Jaegher, and especially Johan van Benthem, Michael Franke and the reviewers for discussion, comments, and suggestions. We also thank Samson de Jager for correcting our English.

of conventional meaning – that conversational implicatures should be accounted for by making use of game theoretic tools. Indeed, in this paper we will propose that a conversational implicature results from an equilibrium outcome of a conversational game between speaker and hearer.

The game theoretic view of implicatures we would like to put forward differs remarkably from the standard analysis associated most with authors like Gazdar, Horn, and Levinson. According to the standard analysis, conversational implicatures depend very directly on the Gricean maxims of conversation – the maxims of *quality*, *quantity*, *relevance*, and *manner* – that specify what participants have to do in order to satisfy Grice's *cooperative principle*. Over the years many phenomena have been explained in terms of the Gricean maxims of conversation. These maxims – best thought of as rules of thumb – are stated in a very informal way, but most work in formal pragmatics has concentrated itself on Grice's two submaxims of *quantity*:

1. Make your contribution as informative as is required
   (for the current purposes of the exchange).

2. Do not make your contribution more informative than is required.

Gricean implicatures based on the first submaxim (sometimes called Q_1-implicatures) are typically analyzed as a rule of negation as failure in the message: from the fact that the speaker didn't say A for a certain class of propositions A, the interpreter infers $\neg A$ (or at least that the speaker doesn't know A). Scalar implicatures as discussed by Horn (1972), Gazdar (1979) and others, from, for instance, '$A \vee B$' and 'John has two children' to 'not $(A \wedge B)$' and 'John doesn't have *more than* two children', respectively, are the best known implicatures of this kind.

The *second* submaxim of quantity is seen as the driving force behind other kinds of pragmatic inferences: those to the most stereotypical interpretation. For example, we normally interpret 'John killed his secretary' as meaning that John murdered his secretary in a stereotypical way, i.e. on purpose and by knife or pistol. For obvious reasons, we will call the latter class of inferences Q_2-implicatures.[1]

A third type of implicature is normally related to Grice's maxim of Manner: "Be perspicuous". More specifically, to his first submaxim "avoid obscurity of expression" and his fourth "avoid prolixity". The idea is that because unmarked expressions give rise to stereotypical interpretations via Q_2-implicatures, if a marked expression is used it is suggested that the stereotypical interpretation should be avoided. For this reason it is argued that the complex 'John caused his secretary to die' rules out that John simply shot his secretary, and suggests that the latter died by an overdose of (secretarial) work. The inference from marked expressions to marked meanings is sometimes called an M-implicature (Levinson, 2000), and the dual pattern described by Q_2- and M-implicatures is dubbed by Horn (1984) the *division of pragmatic labor*.

According to the standard analysis, the three types of implicatures mentioned above are thought of as *generalized* conversational implicatures (GCIs) triggered by specific lexical items. For Q_1-implicatures this means that if the two lexical expressions S(trong) and W(eak) form a scale, $\langle$S, W$\rangle$, a (non-complex) sentence in which the *weaker* expression W occurs will always trigger the implicature that the corresponding *stronger* sentence where S is substituted for W is not true. In some contexts, however, this will give rise to wrong predictions. This problem is usually discussed

[1] What we denote by Q_2-implicatures are called R-implicatures by Horn (1984), and I-implicatures by Levinson (2000).

for *numeral* expressions. Based on the assumption that numerals get an *at least* interpretation, Horn (1972) assumes that they form scales like $\langle ..., \textit{four, three, two}, ... \rangle$. However, the existence of such a scale would falsely predict that A's answer to Q's question implicates that John doesn't have more than two children (cf Kempson, 1986).

(1) Q: Who has two children? A: John has two children.

The *at least* interpretation of numerals has been widely disputed (e.g. Carston, 1998), however. But the phenomenon is not restricted to numerals. Due to the ⟨and, or⟩ scale, it is standardly assumed that (2b) is a Q_1-implicature of (2a):

(2) a. John or Mary came to the party.

b. John or Mary came to the party, but not both.

However, this inference is not always allowed. In particular this is not the case when (2a) is given as an affirmative answer to the following *yes/no*-question:

(3) Did (at least) John or Mary came to the party?

These examples seem to suggest that Q_1-implicatures are, after all, dependent on the conversational situation, in particular, on the question being asked.

Proponents of *generalized* conversational implicatures argue that at least for (1) and (2a), in such *particular* conversational situations, the *generalized* conversational implicatures that John does *not* have *more than* two children and (2b), that John and Mary didn't both come, are *cancelled* for reasons of relevance: the semantic meanings of the answers are already *informative enough* for the purpose of the conversation. Thus, it is claimed by Gazdar and Levinson that Q_1-implicatures are a kind of *default* inference: always triggered, but possibly cancelled.[2] But this makes one suspicious: why should we even *trigger* implicatures for reasons of informativity to be cancelled later for reasons of relevance? Everything else being equal, wouldn't it be preferred to have a theory that can do without cancelation and make the triggering of implicatures dependent on what is known and at stake in the context of use? Grice (1967) called implicatures that crucially depend on context *particularized* conversational ones. The view that Q_1-implicatures depend on what is known and taken to be relevant by speaker and hearer is known as the *Context-Driven* view of Q_1-implicatures.[3]

Both the standard analysis and the Context-Driven view can account for the fact that in (1), and (2a) in the context of question (3), the Q_1-implicature does not arise. However, the different accounts predict different psychological reasoning being involved. Recently, a number of people have tried to decide between the two contrasting approaches of Q_1-implicatures by looking at this psychological reasoning.

Psycholinguistic evidence of at least two different types suggests – though certainly not conclusively[4] – that Q_1-implicatures should not be accounted for as default inferences as proposed by Gazdar, Horn, Levinson, and others. The first type of evidence is that children below age four don't infer standard scalar implicatures (cf. Noveck, 2001; Papafragou & Mussolino, 2003). Noveck (2001), for instance finds

[2] In earlier work, Horn seems to adopt this position as well, but he is less enthusiastic about the view that implicatures are such default inferences in later work (e.g Horn, 2004).

[3] Proponents of the Context-Driven view of Q_1-implicatures include Hirschberg (1985), Carston (1998), and Van Rooij & Schulz (2004).

[4] See Chierchia et al (2001) and Storto & Tannenhaus (2004) for (very) different opinions.

that in contrast to adults, children treat 'Some elephants have trunks' as not being false or misleading if all (shown) elephants have trumps. This evidence is taken to be in favor of the Context-Driven view of Q_1-implicatures, because other well-known experiments (such as 'False belief'-tasks) strongly suggest that such children do not yet have standard folk-psychological abilities, which are exactly the abilities needed to account for the implicature according to proponents of the Context-Driven view.

The first type of evidence in favor of Q_1-implicatures as particularized ones is very indirect. Recently, several people have provided more direct psychological evidence in favor of the Context-Driven view. Recall that in contrast to the Context-Driven view, the standard analysis predicts that in (1), for example, a (potential) implicature is triggered that has to be cancelled later. This indicates that the two analyses predict a difference in reading time for the trigger phrase: the standard analysis predicts that in the context of question (3) the reading time of (2a) takes longer than when the latter was uttered in the context of the question what was in the box. Looking at the reading times of expressions in different types of contexts, Noveck & Posada (2003), Bott & Noveck (2004), and Breheny, Katsos & Williams (2006) consistently find that they confirm the Context-Driven view.

According to Levinson (2000), also Q_2-implicatures are *generalized* ones. The idea is that the semantic meaning of a sentences like 'John killed his secretary' leaves open how John killed his secretary, and whether it was on purpose or not. This should be inferred from contextual, or world knowledge. This all sounds very natural, but why should it be a *generalized* implicature, instead of a *particularized* one?[5] Because what counts as stereotypical killing depends on time and place (compare the Wild West with Italy during the Roman Empire), it seems most natural to assume that a Q_2-implicature depends on the common (though perhaps very idiosyncratic) background assumptions of speaker and hearer. But this suggests that one should also treat inferences to stereotypical interpretations as *particularized* conversational implicatures. This reasoning caries over to M-implicatures, because they depend on what is mutually understood to be stereotypical as well.

We have argued against the assumption that standard pragmatic inferences should be treated as *rule-governed*, *generalized* conversational implicatures. Instead, Q_1, Q_2, and M-implicatures should all be treated as *particularized* conversational implicatures, dependent on the preferences and beliefs speaker and hearer have in the particular conversational context. This suggest (i) that all these implicatures should be accounted for similarly, but also (ii) that they have a lot in common with the prototypical particularized conversational implicatures: the ones dependent on Grice's maxim of *Relevance*: the requirement of the speaker to be relevant. In the following section we will use decision theory to define a precise notion of relevance, and see how we can use it to account for some implicatures. Although the analysis is quite successful to accounting for standard Q_1-implicatures and some relevance implicatures, the proposed analysis is not general enough to account for all relevance implicatures and fails to account for Q_2-implicatures. Afterwards, we will provide a more general game theoretic analysis of conversational implicatures which is independent of Grice's specific maxims, and show by example how it can treat successfully all types of implicatures mentioned in this introduction.

[5] See also Geurts (1998).

2 Maximizing relevance

The standard analysis of Q_1-implicatures assumes that the speaker says as much as he can. The hearer can conclude from this that all those alternative expressions that the speaker could have used can be taken to be false if they are *more informative*. It is very straightforward to account for this type of pragmatic inference by means of the following pragmatic interpretation rule:

$$Prag_1(f) = \{w \in [\![f]\!] | \forall f' \in Alt(f) : w \in [\![f']\!] \rightarrow f \models f'\}. \tag{2.1}$$

In this rule, $Alt(f)$ denotes the set of alternative expressions to f that the speaker could have used, $[\![\cdot]\!]$ is a function which assigns to each expression f its semantic interpretation $[\![f]\!]$, a set of worlds, while $f \models f'$ denotes the fact that f *entails* f' (meaning that $[\![f]\!] \subseteq [\![f']\!]$). Obviously, if the speaker asserts '$A \vee B$', the hearer can conclude via the pragmatic interpretation rule $Prag_1$ that the stronger '$A \wedge B$' is false (if '$A \wedge B$' is taken to be an alternative to '$A \vee B$').

Although $Prag_1$ accounts for a wide range of conversational implicatures due to Grice's first submaxim of Quantity, there are quite a number of shortcomings of this interpretation rule. For instance, it puts heavy constraints on the alternative expressions (for '$A \vee B$' to be interpretable at all, 'A' is not allowed to be an alternative), and it assumes the speaker to be *knowledgeable* about which of the alternative expressions are true. But even if we ignore these problems, the above pragmatic interpretation rule has an obvious shortcoming: it ignores the purpose of the conversation. This shortcoming can be illustrated by the following standard conversation taking place at the Damrak in Amsterdam between Italian tourist Ann and Dutchman Bob:

(4) Ann: Where can I buy Italian wine?
Bob: At the central station. (f_1) / At the Bijenkorf. (f_2)
Bob: At the central station and at the Bijenkorf. ($f_1 \wedge f_2$)

Intuitively, Bob's answer f_1 does not rule out that one can also buy Italian wine at other places in Amsterdam, for instance at the Bijenkorf. However, if we take 'At the central station and at the Bijenkorf' ('$f_1 \wedge f_2$') to be an alternative to Bob's answer, interpreting by $Prag_1$ wrongly predicts that one cannot buy Italian wine at the Bijenkorf. Arguably, what is ignored by $Prag_1$ is Ann's purpose of asking the question: Ann wants to buy a bottle of Italian wine, and she just needs to know *some* place where one can be bought (Bob's answer is known as a 'mention-some' answer). Knowing all places in Amsterdam is *irrelevant* to this goal.

A natural way to improve on interpretation rule $Prag_1$ – and in line with Grice's rider to his first submaxim of Quantity – is to assume that the speaker said as much as possible as far as this is *relevant* to the hearer. In order to make this suggestion more precise, it has been proposed (Parikh, 2001; Van Rooij, 2003) that the relevance of a proposition is defined in terms of the extent to which it solves the hearer's *decision problem*. We will first discuss what a decision problem is and one simple way to define a relevance ordering between propositions making use of such a decision problem. Afterwards, we suggest how one can account for the mention-some answers.

Let Ω be the set of all possible states of the worlds. For simplicity we restrict our attention to situations with countable many possibilities, i.e. to countable Ωs. We represent an agent's expectations about the world by a probability distribution over Ω, i.e. a real valued function $P : \Omega \longrightarrow \mathbf{R}$ with the following properties: (1) $P(v) \geq 0$ for all $v \in \Omega$ and (2) the sum of all $P(v)$ equals 1. For sets $A \subseteq \Omega$ we

set $P(A) = \sum_{v \in A} P(v)$. Hence $P(\Omega) = 1$. We represent an agent's preferences over outcomes of actions by a real valued utility function over action-world pairs. We collect these elements in the following structure:

Definition 2.1 *A* decision problem *is a triple* $\langle (\Omega, P), \mathcal{A}, U \rangle$ *such that* (Ω, P) *is a countable probability space,* $\mathcal{A}$ *a finite, non–empty set and* $U : \mathcal{A} \times \Omega \longrightarrow \mathbf{R}$ *a function.* $\mathcal{A}$ *is called the* action *set, and its elements* actions. *U is called a* payoff *or* utility *function.*

Let us now assume that our agent, Ann, faces a *decision problem*, i.e. she wonders which of the alternative actions in $\mathcal{A}$ she should choose. It is standard to assume that rational agents try to maximize their expected utilities. The *expected utility* of an action a is defined by:

$$EU(a) = \sum_{v \in \Omega} P(v) \times U(a, v). \tag{2.2}$$

The expected utility of performing an action might change if our agent Ann learns new information. To determine this change of expected utility, we first have to know how learning new information will effect Ann's beliefs. In probability theory the effect of learning a proposition A is modeled by *conditional probabilities*. Let H be any proposition, e.g. the proposition that one sells Italian wine at the station. H collects all possible worlds in Ω where this sentence is true. Let C be some other proposition, e.g. the answer given by Bob. Then, the probability of H *given* C, written $P(H|C)$, is defined by:

$$P(H|C) := P(H \cap C)/P(C). \tag{2.3}$$

This is only well-defined if $P(C) \neq 0$. In terms of this conditional probability function, we can now define the *expected utility of action a after learning C* by:[6]

$$EU(a|C) = \sum_{v \in \Omega} P(v|C) \times U(a, v). \tag{2.4}$$

In terms of this notion, we can determine the *utility value* of the information C, $UV(C)$, as follows:

$$UV(C) \quad = \quad \max_i EU(a_i|C). \tag{2.5}$$

Now that we know how to determine the utility of a proposition (for a hearer) we can define a relevance ordering on these propositions in terms of their utility values: $A \geq_R B$ iff $UV(A) \geq UV(B)$. Because each expression denotes a unique proposition, the ordering relation can be extended to expressions: $f \geq_R f'$ iff $[\![f]\!] \geq_R [\![f']\!]$.

Now one might propose that the speaker should assert something true that has the highest utility value. This would mean that if F is the set of alternative assertions the speaker could make, and we assume that Ann faces a mutually known decision problem, f should pragmatically be interpreted as follows:

$$Prag_2(f) = \{w \in [\![f]\!] | \forall f' \in F : w \in [\![f']\!] \rightarrow f \geq_R f'\}. \tag{2.6}$$

Notice that if $F = Alt(f)$, $Prag_2(f)$ is just like $Prag_1(f)$ except that entailment is replaced by $\geq_R$.

[6] Where $P(v|C)$ is short for $P(\{v\}|C)$.

To illustrate how $Prag_2$ works for our above mention-some example, suppose that we have four relevant worlds, $\Omega = \{w_0, w_1, w_2, w_3\}$, where the three assertions: f_1, f_2, and $f_1 \wedge f_2$ have the following semantic meanings: $[\![f_1]\!] = \{w_1, w_3\}$, $[\![f_2]\!] = \{w_2, w_3\}$, and $[\![f_1 \wedge f_2]\!] = \{w_3\}$. Suppose that the decision problem contains only two actions: a, the action of walking to the central station, and b, the action of walking to the Bijenkorf, and that the utility function is defined as follows: $\forall w \in \Omega : U(a, w) = 1$ if $w \in [\![f_1]\!]$, 0 otherwise, and $U(b, w) = 1$, if $w \in [\![f_2]\!]$, 0 otherwise. If we assume for simplicity that $P(X) = \frac{|X|}{|\Omega|} = \frac{|X|}{4}$, for $X \subseteq \Omega$, it follows that $EU(a|[\![f_1]\!]) = ((\frac{1}{2} \times U(a, w_1)) + (\frac{1}{2} \times U(a, w_3))) = 1 = EU(a|[\![f_1 \wedge f_2]\!]) = U(a, w_3)$, and thus that $UV([\![f_1]\!]) = UV([\![f_1 \wedge f_2]\!]) = 1$. But this means that even though $f_1 \wedge f_2$ is more informative than f_1, it is not more relevant, or useful. Hence, if we interpret f_1 pragmatically by $Prag_2$, one cannot conclude anymore from f_1 that the stronger $f_1 \wedge f_2$ is false.

Unfortunately, there are a number of implicatures we cannot account for in terms of relevance maximization, i.e., interpretation rule $Prag_2$. First of all, it's unclear how Q_2-implicatures should be treated. More disturbingly, perhaps, it doesn't predict correctly in case of the implicature explicitly discussed by Grice (1989) in his William James lectures. In this example, Ann is standing by an obviously immobilized car and is approached by Bob, after which the following exchange takes place

(5) Ann: I am out of petrol.
Bob: There is a garage round the corner. (G)
+> there is petrol available at the garage (H)

Grice notes that because Bob's remark can only be relevant in case the garage is open, H, Ann can conclude that this is something conversationally implicated by Bob. As suggested above, this is not predicted by interpretation rule $Prag_2$. Relevance maximization predicts that everything is false what would have been more relevant and was not explicitly stated. H was not stated and would have been relevant, hence it is predicted to be false by the pragmatic interpretation rule $Prag_2$. What is wrong, or so we will argue, is that we should not (just) look at maximizing utility from the *hearer's* point of view, but (also) from the perspective of the *speaker*: what counts is the speaker's utility of the action chosen by the hearer after she updated her beliefs with the semantic meaning of the answer. To account for this, we will make use to game theoretic techniques.

3 A game theoretic analysis

3.1 Representing the situation

In the previous section we have referred to Ann's decision problem to determine the relevance of new information, and tried to calculate the 'pragmatic' meaning of the answer in terms of that. Unfortunately, we saw that this sometimes gives rise to wrong predictions. In this section we will argue that to solve this problem, we should embed Ann's decision problem into a larger interactive setting that involves not only a decision problem of Ann's, but also one of Bob's. Bob's decision problem is the problem which answer he should provide such that Ann chooses the optimal action. We will argue that Ann can figure out what is conversationally implicated by the answer, or more generally by Bob's assertion, if she assumes that Bob gave the answer which gives the highest expected utility to *himself*. But for Bob to calculate

his expected utility of an answer, he has to make certain assumptions about how Ann will (initially) update her beliefs as a response to the answer. The crucial proposal we will make in this paper (following Benz, 2006), is that this initial update depends only on the *semantic* meaning of the answer. In this section we (slightly) extend Benz's (2006) analysis and show how to represent dialogues as two person games with complete coordination, and demonstrate how speaker Bob can determine the optimal answers/assertions by means of backwards induction. Afterwards, we show how – or under which circumstances – this predicts the intuitively correct implicatures triggered by assertions.

To a game theorists, our use of backwards induction will be surprising, given that it is not the case that the participants of the game have perfect information. We will make use of very special assumptions about how Ann interprets messages and chooses her actions in response, however, which allows us to make use of backwards induction after all. Intuitively, these special assumptions will implement the idea that Ann can trustfully presuppose that Bob will give a maximally cooperative answer. In section 5 we will compare our game theoretical model of communication with a more standard one using signaling games, and show that the behavior we predict by backwards induction in our model corresponds with a particular type of solution in the standard approach.

Instead of Ann and Bob, we will in this section talk about the *inquirer*, I, and the answering *expert*, E. The inquirer has a decision problem, $\langle(\Omega, P_I), \mathcal{A}, U_I\rangle$ and we will assume for simplicity that this problem is common knowledge after she has asked her question. In order to get a model for the questioning and answering situation we have to add a representation for the answering expert's situation. We only add a probability distribution P_E that represents his expectations about the world:

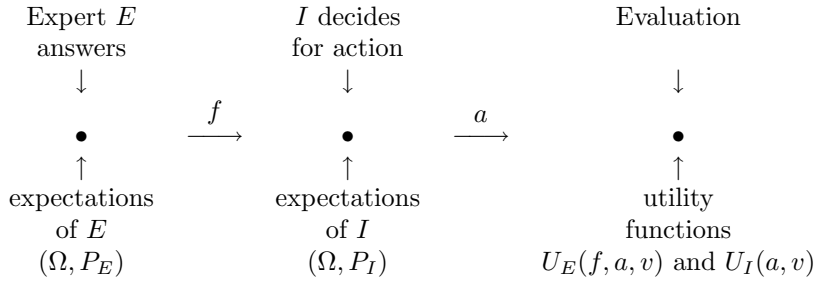
Definition 3.1 *A* support problem *is an eight–tuple* $\langle \Omega, P_E, P_I, F, \mathcal{A}, U_E, U_I, [\![\cdot]\!]\rangle$ *such that* (Ω, P_E) *and* (Ω, P_I) *are countable probability spaces,* $[\![\cdot]\!]$ *is the semantic interpretation function, and* $\langle(\Omega, P_I), \mathcal{A}, U_I\rangle$ *and* $\langle(\Omega, P_E), F, U_E\rangle$ *are decision problems.*[7] *In contrast to* U_I*,* U_E *will depend also on the message being used. We define* $U_E(f, a, w)$ *as* $U_I(a, w) - C(f)$*, where* $C(f)$ *measures the cost of sending* f*. We call a support problem* well–behaved *if (1) for all* $B \subseteq \Omega$*:* $P_I(B) = 1 \Rightarrow P_E(B) = 1$ *and (2) for* $x = I, E$ *and all* $a \in \mathcal{A}$*:* $\sum_{v\in\Omega} P_x(v) \times |U_x(\cdot, a, v)| < \infty$*.*

We assume that all the elements of a support problem are common knowledge, except for P_E. As far as cost function C is concerned, we follow Blume, Kim & Sobel (1993) and assume that either $C(f) = 0$, or $C(f)$ is nominal, i.e. small relative to other payoffs. The first condition for well-behavedness is included in order to make sure that E's answers cannot contradict I's beliefs. It implies that for any $f \in F$ and set $B \subseteq \Omega$: $P_E([\![f]\!]) = 1 \Rightarrow P_I([\![f]\!]) > 0$ and $P_I([\![f]\!]|B) = 1 \,\&\, P_E(B) = 1 \Rightarrow P_E([\![f]\!]) = 1$. The second condition in the definition is there in order to keep the mathematics simple.

A support problem represents just the fixed static parameters of the answering situation. A crucial assumption we make is that I's decision does not depend on what she believes that E believes. Hence her epistemic state (Ω, P_I) represents just her expectations about the actual world. E's task is to provide information that is optimally suited to support I in her decision problem. Hence, E faces a decision problem himself, where his actions are the possible answers. The utilities of the

[7] We leave it underdetermined here, but the set F should most naturally be thought of as the set of alternative answers to the question 'corresponding' to I's decision problem.

answers depend on how they influence I's final choice. We look at the dependencies in more detail. We find two successive decision problems:



We assume that the answering expert E is fully cooperative and wants to maximize I's final success. If we ignore E's costs of sending the messages, E's payoff is identical with I's (our representation of the *Cooperative Principle*). E has to choose his answer in such a way that it optimally contributes towards I's decision. Due to our assumption that I's information is mutually known, E is able to calculate how I will decide. Hence, we represent the decision process as a sequential two-person game with (almost) complete coordination of preferences. We find a solution, i.e. optimal assertions and choices of actions by calculating backwards from the final outcomes. The following model will be worked out concentrating on *ideal* dialogues.

3.2 Calculating optimal answers by backwards induction

I's decision situation First we have to consider the final decision problem of I. In the previous section we have determined the *expected utility after learning f* by:

$$EU_I(a, f) = EU_I(a|[\![f]\!]) = \sum_{v \in \Omega} P(v|[\![f]\!]) \times U_I(a, v).$$

If the decision maker I tries to maximize expected utilities by her choice, it follows that she will only choose actions that belong to $argmax_a EU(a, f) = \{a \in \mathcal{A} \mid EU_I(a, f)\ \textit{is maximal}\}$. Sometimes we assume in addition that I has always a preference for one action over the other, or that there is a mutually known rule that tells I which action to choose if this set has more than one element. In this case we can write a_f for this unique element. In short, we assume that the function $f \mapsto a_f$, for $P_I([\![f]\!]) > 0$, is known to E.

E's decision situation According to our assumption, questioning and answering is a game of (almost) complete coordination (Principle of Cooperation). We have implemented this assumption by taking E's utility function U_E to be identical with I's utility function U_I minus the nominal cost of sending the message. We use a slight variant of definition (2.2) for calculating the expected utility of an answer $f \in F$. With a_f as defined above we define:

$$EU_E(f) := \sum_{v \in \Omega} P_E(v) \times U_E(f, a_f, v). \tag{3.7}$$

Notice that on our assumption that $C(f)$ is at most nominal, in order to maximize his own benefit, E has to choose an answer such that it induces I to take an action

that maximizes their common payoff. We add here a further Gricean maxim, the *Maxim of Quality*. We call an answer f *admissible* if $P_E([\![f]\!]) = 1$. The Maxim of Quality is represented by the assumption that the expert E only gives admissible answers. This means that he believes them to be *true*. For a support problem $\sigma = \langle \Omega, P_E, P_I, F, \mathcal{A}, U_I, U_E, [\![\cdot]\!] \rangle$ we set:[8]

$$Adm_\sigma := \{f \in F : P_E([\![f]\!]) = 1\}. \tag{3.8}$$

Hence, the set of optimal answers for σ is given by the set of admissible answers that have the highest expected utility:

$$\mathrm{Op}_\sigma = \{f \in Adm_\sigma \mid \forall f' \in Adm_\sigma : EU_E(f) \geq EU_E(f')\}. \tag{3.9}$$

Assuming that the expert is making an optimal assertion, the inquirer can conclude from E's assertion f that she is in a support problem σ where it holds that $f \in \mathrm{Op}_\sigma$. Because, by assumption, she knows already E's utility function, I can learn a lot about the information E has about the actual world.

Let us assume that E knows for sure which action $a \in \mathcal{A}$ is optimal. This means that there must be some action a such that a is optimal in all worlds which are possible to E. Let $O(a)$ denote the set of worlds where action a has highest utility for E and I (remember that the cost of messages is at most nominal):

$$O(a) = \{w \in \Omega \mid \forall b \in \mathcal{A} : U_E(\cdot, a, w) \geq U_E(\cdot, b, w)\}. \tag{3.10}$$

For E to know which action is optimal, it has to be the case that there is an action a such that he assigns to proposition $O(a)$ probability 1:

$$\exists a \in \mathcal{A} : P_E(O(a)) = 1. \tag{3.11}$$

Let's assume that E asserted f, and that this assertion was actually optimal for E, and was assumed by I to be optimal, i.e. that $f \in \mathrm{Op}_\sigma$. Thus,

$$P_E([\![f]\!]) = 1 \wedge \forall f' : (P_E([\![f']\!]) = 1 \rightarrow EU_E(f) \geq EU_E(f')). \tag{3.12}$$

It follows with (3.11) and (3.12) that E knows he is in a world where a_f is the action, or one of the actions, which has the highest utility:

$$P_E(O(a_f)) = 1. \tag{3.13}$$

From this I can infer that the actual world w is such a world as well:

$$w \in O(a_f). \tag{3.14}$$

What is interesting about this inference is that although E determines via backwards induction what he should assert by making use of naive Bayesian updating on the inquirer's side, E in fact realizes that on the basis of this assumption I will update her beliefs via a more sophisticated method than conditionalization. We will see next that in this way we predict mention-some readings of answers and Grice's relevance implicature in the circumstances described in section 2.

[8] We assume that there is an $f \in F : [\![f]\!] = \{w \in \Omega | P_E(w) \neq 0\}$.

3.3 Mention-some answers and Relevance implicatures

Let us consider the mention-some question in (4) again in the situation as described in section 2.

(6) I: Where can I buy Italian wine?

E: At the central station. (f_1) / At the Bijenkorf. (f_2)

E: At the central station and at the Bijenkorf. ($f_1 \wedge f_2$)

Recall that the answers (f_1) and (f_2) are called *mention-some* answers. The answer ($f_1 \wedge f_2$) is more informative than both of these.

Let us denote by a and b the actions of going to the station and going to the Bijenkorf, respectively. There may be other actions too. Let $[\![f_1]\!] \subseteq \Omega$ be the set of worlds where one can buy Italian wine at the station, and $[\![f_2]\!] \subseteq \Omega$ where one can buy Italian wine at the Bijenkorf. For every possible action $c \in \mathcal{A}$ the utility value is either 1 (success) or 0 (failure); especially we assume that $U_I(a, v) = 1$ iff $v \in [\![f_1]\!]$, else $U_I(a, v) = 0$; $U_I(b, v) = 1$ iff $v \in [\![f_2]\!]$, else $U_I(b, v) = 0$.

In section 2 we showed already by way of an example that $UV_I([\![f_1]\!]) = 1 = UV_I([\![f_1 \wedge f_2]\!])$ and it similarly holds that $UV_I([\![f_2]\!]) = 1 = UV_I([\![f_1 \wedge f_2]\!])$. This shows that for the inquirer it doesn't matter which information she receives, as long as it is true. Thus, all the answers are equally useful with respect to the conveyed information and the inquirer's goals. What we have to show now, however, is that all answers are equally optimal for the answering *expert*. We will show that $EU_E(f_1) = EU_E(f_2) = EU_E(f_1 \wedge f_2) = 1$ if f_1, f_2 and $f_1 \wedge f_2$ are admissible answers, and thus known to be true by the expert.

We start with answer f_1: If E knows that f_1 is true, then f_1 is an optimal answer. We assume that f_1 is a costless message, meaning that $U_E(f_1, a_{f_1}, v) = U_I(a_{f_1}, v)$ for all $v \in \Omega$. If learning $[\![f_1]\!]$ induces I to choose action a, i.e. if $a_{f_1} = a =$ going to the central station, then the proof is very simple:

$$EU_E(f_1) = \sum_{v \in \Omega} P_E(v) \times U_E(f_1, a_{f_1}, v) = \sum_{v \in [\![f_1]\!]} P_E(v) \times U_I(a, v) = 1.$$

Clearly, no other answer could yield a higher payoff. If we want to prove the claim in full generality, i.e. for all cases, may they be as complicated as they can be as long as our previously formulated restrictions hold, then we need some more calculation. We first note the following fact: Let's assume that I chooses after learning $[\![f_1]\!]$ an act c different from a, i.e. $a_{f_1} = c \neq a$. Then let $O(c)$ denote the set where action c is successful, i.e. $O(c) = \{v \in \Omega \mid U_I(c, v) = 1\}$. Then either (i) $P_E(O(c)) = 1$ or (ii) $P_E(O(c)) < 1$. In the first case (i) it follows again that $EU_E(f_1) = \sum_{v \in \Omega} P_E(v) \times U_E(f_1, c, v) = 1$, and our claim is proven. Case (ii) leads to a contradiction by the well–behavedness condition in Definition 3.1: If I chooses c, then $EU_I(c|[\![f_1]\!]) = \max_{c' \in \mathcal{A}} EU_I(c'|[\![f_1]\!]) = EU_I(a|[\![f_1]\!]) = 1$; hence $P_I(O(c)|[\![f_1]\!]) = 1$, and therefore $P_E(O(c)) = 1$ by well-behavedness, in contradiction to (ii). It follows that only (i) is possible.

In the same way it follows that f_2 is optimal if E knows that f_2. If all messages are costless, the same result follows for any stronger answer, including $f_1 \wedge f_2$, $f_1 \wedge \neg f_2$, or $\neg f_1 \wedge f_2$. This shows that their expected utilities are all equal as long as they are admissible answers. Hence, all these answers are equally good and E can freely choose between them. But this means that the pragmatic interpretation of f_1 is the

same as its semantic meaning, and thus that it will receive a mention-some reading. Notice that if $C(f_1 \wedge \neg f_2) > C(f_1)$, asserting f_1 would be more optimal than asserting $f_1 \wedge \neg f_2$, even if the latter statement is true and more informative. This reasoning very much satisfies Grice's second submaxim of Quantity as discussed in section 1.

We have seen in section 2 that we can already predict that f_1 receives a mention-some reading by interpretation rule $Prag_2$, but that this rule makes the wrong prediction for Grice's (1967) example repeated below:

(7) A: I am out of petrol.

B: There is a garage around the corner. (G)

Again, Grice suggests that because B's remark can only be relevant in case the garage is open, $H = \{w_2\}$, A can conclude that this is something conversationally implicated by B. Let us assume that a denotes the action of having a look around the corner, while b denotes the action of doing nothing. We assume here that going to the garage without getting petrol is more costly than doing nothing in world w_1 where the garage doesn't have petrol. This can be represented by e.g. the following payoff function: $U_I(a, w_1) = -1$ and $U_I(a, w_2) = 10$ and $U_I(b, w_1) = U_I(b, w_2) = 0$. We assume that I's expectations are such that learning that (G) there is a garage around the corner will induce her to do a.[9] From this it follows that the real world w must be such that $U_I(a, w) \geq U_I(b, w)$. This can only be the case if $w = w_2$. Hence, the fact that G has been answered *implies* that the garage is open and offers petrol.

4 The standard implicatures

4.1 Calculating implicatures

We have argued that an informed speaker, i.e. the expert, can and should use backwards induction to determine which answer he should give. Notice that by our use of backwards induction, the informed speaker assumes that I will perform that act which has the highest expected utility after she has updated her beliefs by standard Bayesian conditionalization with the *semantic* meaning of the answer. This doesn't mean, however, that the hearer I interprets the answer simply at face value: On the assumption that E is informed of I's decision problem and chooses his answer by making use of backwards induction, hearer I can conclude more from the answer than just its standard semantic meaning. Before we will discuss the interpretation rule that is associated with the speaker's rule of backwards induction, let us first discuss some other, and perhaps simpler, combinations of speaker and hearer strategies.

It is well-known (e.g. Levinson, 2000) that we can look at the Gricean maxims both from the speaker's and from the hearer's perspective. The quality maxim, for instance, can be thought of as a requirement for the speaker to speak the truth, and if the maxim is obeyed the hearer can conclude that what the speaker says is (believed to be) true. The same is true for the first sub-maxim of quantity, which basically demands speakers to provide all information they know as far as this is relevant to

[9] This model may seem to be somewhat artificial. In a realistic model we have to assume that there are many different places where it might be possible that petrol is available. This means that I has to choose between a larger number of actions. In such a scenario it becomes very natural to assume that only learning G will induce her to do a. But in order to keep the model simple, we consider only a situation where I has to choose between doing nothing and going to that specific garage.

the current topic of conversation. If it is known that the speaker obeys this maxim, the hearer can conclude something on top of what is explicitly said by the speaker.

Let us now return to the case where the speaker, or expert, determines what to say by means of backwards induction. We can think of a speaker's strategy as a rule that determines for each support problem what the speaker should say. In this case it is the function that says that E should utter f, if the action which has the highest expected utility for I after she learns that f is true is indeed among the best actions according to E.

In the previous sub-section we saw that an answer/assertion must be an element of Op_σ for any support problem $\sigma = \langle \Omega, P_E, P_I, F, \mathcal{A}, U_E, U_I, [\![\cdot]\!] \rangle$. But this means that hearer I can conclude something about the speaker's E's decision problem: he must have been in a decision problem where the answer was an optimal assertion. If we assume that the alternative actions F that the speaker can perform and his utility function U_E are common knowledge, this means that I can learn something about the beliefs of E. Remember that if E answered f, then the inquirer I knows that $E's$ answer is optimal, i.e. that $f \in \mathrm{Op}_\sigma$, hence that:

$$P_E([\![f]\!]) = 1 \ \wedge \ \forall f' \in F : (P_E([\![f']\!]) = 1 \rightarrow EU_E(f) \geq EU_E(f')).$$

If we assume that it is common knowledge between E and I that E has complete knowledge of all relevant facts, then it follows that for the actual world w it holds that $P_E(w) = 1$. In that case, if it is assumed that E makes an optimal assertion, I can conclude from E's assertion f that E and I must be in one of the following worlds:

$$Prag_3(f) = \{w \in [\![f]\!] | \forall f' \in Adm_\sigma : U_E(f, a_f, w) \geq U_E(f', a_{f'}, w)\}. \tag{4.15}$$

As it turns out, in terms of pragmatic interpretation rule $Prag_3$ we can account for all the implicatures discussed in the introduction. What we need to assume is that all that matters for the hearer is to learn what the real world is. To implement this, we can assume that the set of actions $\mathcal{A}$ corresponds 1-1 with the set of worlds, i.e. that there exists a 1-1 function, τ, between actions and worlds. This means that I's utility function is just $U_I(a, w) = 1$ if $\tau(a) = w, 0$ otherwise. But then expected utility reduces to probability: $EU_I(a, f) = P_I(w|[\![f]\!])$, if $w = \tau(a)$, and maximizing expected utility reduces thus to maximizing probability. Perhaps in contrast to other examples, there seems no reason for I to prefer world w to world v if both have the same conditional probability given f. To account for that we assume that $a_f = argmax_a EU_I(a, f)$ is just the set of maximally likely $[\![f]\!]$-worlds: $\{v \in [\![f]\!] : v \in argmax_w P_I(w|[\![f]\!])\}$.

Let us now assume that it is commonly known that speaker E knows in which (relevantly different) state he is, and thus that I can, in general, infer from E's assertion f that one of the worlds in $Prag_3(f)$ is the actual one. Given that $a_f = argmax_w P_I(w|[\![f]\!])$, and assuming for simplicity that all messages in F are equally complex, this suggests that $U_E(f, a_f, w)$ measures the chance that I interprets the message in the correct way. Because I takes by construction all worlds in a_f to be equally likely, $U_E(f, a_f, w)$ reduces to $P_I(w|a_f) = P_I(w|argmax_w P(w|[\![f]\!]))$. On this assumption, (4.15) reduces to the following interpretation rule:

$$Prag_3(f) = \{w \in [\![f]\!] | \forall f' \in Adm_\sigma : P_I(w|a_f) \geq P_I(w|a_{f'})\}, \tag{4.16}$$

which, in turn, simplifies to

$$Prag_3(f) = \{w \in [\![f]\!] | \forall f' \in F : P_I(w|a_f) \geq P_I(w|a_{f'})\}. \tag{4.17}$$

4.2 Quantity$_1$ implicatures

In this section we will argue that the interpretation rule derived in the previous section immediately accounts for Q_1-implicatures if $a_f = [\![f]\!]$, for each $f \in F$. Perhaps we should simply make this assumption, or we can derive it if we assume that all $[\![f]\!]$-worlds are equally likely. In any case, if $a_f = [\![f]\!]$, pragmatic interpretation rule (4.17), and thus (4.15), comes down to the following:[10]

$$Prag_3(f) = \{w \in [\![f]\!] | \forall f' \in F : P_I(w|[\![f]\!]) \geq P_I(w|[\![f']\!])\}. \tag{4.18}$$

Notice that this rule has a lot in common with interpretation rule $Prag_2$ discussed in section 2. The main difference (in case I's expected utility comes down to her probability) is that the world w that *the speaker* knows to be the actual one plays a crucial role in $Prag_3$, but not in interpretation rule $Prag_2$. And intuitively it *should* play a crucial role: the speaker who knows he is in w wants the hearer to reach that conclusion, which is why this fixed world should play a prominent role in I's reasoning.

To illustrate how rule (4.18) accounts for Q_1-implicatures, let us look for simplicity at numerical expressions. Let us assume that $\Omega = \{w_1, w_2, w_3, w_4\}$, where w_i is the world where i children came to the party, and that E can choose between four messages: $F = \{$'one','two','three','four'$\}$ with their standard neo-Gricean 'at least'-interpretations. This means that the meanings of the numeral expressions form an implication chain: [\!['four']\!] $\subset$ [\!['three']\!] $\subset$ [\!['two']\!] $\subset$ [\!['one']\!].[11] By Quality, E has to say something that is true. If the speaker is in world w_4 where 4 children came to the party, he could send all four messages, but if he is in w_1, a world where only 1 child came, he could say only that. What could I conclude from the message if E knows in which world he is in? The following table will help us to see.

$P_I(w\|[\![f]\!])$	w_1	w_2	w_3	w_4
'one'	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
'two'	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
'three'	0	0	$\frac{1}{2}$	$\frac{1}{2}$
'four'	0	0	0	1

From this table we can see that I can conclude via $Prag_3$ that each numerical expression 'n' will pragmatically be strengthened from 'at least n' to mean 'exactly n'. This is so, because if E were in a state $m > n$, there would be an alternative expression which has a higher utility/probability in that state in the sense that there is a higher chance that it is interpreted as intended by E.

4.3 Q_2-implicatures, and Horn's division of pragmatic labor

In this section we will show that $Prag_3$ can also account for inferences to stereotypical interpretation (i.e., Q_2-implicatures) and for *Horn's division of pragmatic labor* –

[10] This interpretation rule has, in fact, been proposed already by Van Rooij (2004a) in the context of Bidirectional Optimality Theory. Jäger's (2006) game theoretic analysis of Q_1-implicatures comes down to this rule as well.

[11] As noted in the introduction, the assumption that numerical expressions have an 'at least' interpretation is highly controversial, and probably even wrong (see also Clark & Grossman, to appear). For the argument it doesn't matter much: one could easily think of other examples where the semantic meanings of the alternative expressions form a linear chain with respect to inference. The scale ⟨and, or⟩ would do as well, just as ⟨all, most, some⟩, if the quantifiers 'all' and 'most' give rise to an existential presupposition.

according to which an (un)marked expression (morphologically complex and less lexicalized) typically gets an (un)marked meaning. To illustrate, consider the following well-known example:

(8) a. John killed the secretary.

b. John caused the secretary to die.

We typically interpret the unmarked (8a) as stereotypical killing, i.e., that John killed the secretary by knife or pistol (a Q_2-implicature), while the marked (8b) is interpreted such that John caused the death of his secretary in a non-standard way, e.g. by an overdose of (secretarial) work. In this section we will see how we can account for this pattern in our framework.

First, remember that via the pragmatic interpretation rule $Prag_3$:

$$Prag_3(f) = \{w \in [\![f]\!] : \forall f' \in Adm_\sigma : U_E(f, a_f, w) \geq U_E(f', a_{f'}, w)\},$$

I can conclude from the use of cost-free f that the actual world is one of the most plausible worlds compatible with f: it is an element of $a_f = argmax_w P(w|[\![f]\!])$. Thus, $Prag_3(f) = \{w \in [\![f]\!] | \forall f' \in F : P_I(w|a_f) \geq P_I(w|a_{f'})\}$. If in contrast to what we did for Q_1-implicatures we now don't assume, or derive, that $a_f = [\![f]\!]$, this means that we can (almost) immediately infer to a stereotypical interpretation, also known as a Q_2-implicature.[12]

Making use of the same utility functions, we can also account for M-implicatures: the fact that marked expressions typically receive a marked interpretation. Suppose we have 2 (types of) worlds, w_1 and w_2, and 3 messages, f_u, f_1 and f_2. Assume that f_u has an underspecified meaning, $[\![f_u]\!] = \{w_1, w_2\}$, while f_1 and f_2 have a specific meaning: $[\![f_1]\!] = \{w_1\}$, and $[\![f_2]\!] = \{w_2\}$. Let us assume, moreover, that $P_I(w_1) = P_I(w_1|[\![f_u]\!]) > P_I(w_2|[\![f_u]\!]) = P_I(w_2)$. As before, E's utility function will be decomposable into I's utility function as defined above and a cost function, C. We assume that $C(f_u) = 0$ and $C(f_1) = C(f_2) > 0$, though small. Assuming that there exists a 1-1 correspondence between worlds and actions, this means that for each message f there is a unique action with the highest expected utility for I, denoted by a_f. Notice that $a_{f_u} = a_{f_1} = w_1 \neq w_2 = a_{f_2}$.

Given this, we can calculate for each message its expected utility for E: $EU_E(f) = \sum_{v \in \Omega} P(v) \times U_E(f, a_f, v)$. If we assume that E knows in which (type of) world he is, $EU_E(f)$ reduces to $U_E(f, a_f, w)$ for the known (type of) world w. I can now interpret E's message again by interpretation rule $Prag_3$. Notice that $Prag_3(f_u) = \{w_1\}$, $Prag_3(f_2) = \{w_2\}$, while $Prag_3(f_1) = \emptyset$. For $Prag_3(f_1)$ to be the empty set, it has to be the case that there is another expression whose semantic meaning includes w_1 and sending it has a higher utility. This means that it has to be the case that $U_E(f_u, a_{f_u}, w_1) > U_E(f_1, a_{f_1}, w_1)$. Although $a_{f_u} = a_{f_1}$, this is the case because $C(f_u) < C(f_1)$. Notice that with the utility function U_E as defined above, the pragmatic interpretation rule $Prag_3$ encodes Horn's division of pragmatic labor: the less expensive message f_u receives the stereotypical interpretation w_1, while the marked interpretation w_2 has to be expressed by a marked message, f_2.

[12] Almost, because if a_f contains more worlds, there might in principle be an alternative expression f' with $a_{f'} \subset a_f$, and thus $P_I(w|a_{f'}) > P_I(w|a_f)$. We have to assume that there are no such alternatives $f' \in F$.

5 Comparison with a signaling game approach

In this section we relate support problems to signaling games of complete coordination. David Lewis (1969) introduced signaling games to study (linguistic) conventions, and extensions of these games have been studied afterwards in economics and biology. In Game Theory textbooks (e.g. Fudenberg & Tirole, 1991; Gibbons, 1992), signaling games are seen as dynamic (or extensive form) Bayesian games. The standard solution concept associated with Bayesian games is that of a *perfect Bayesian equilibrium*. In this section we first show that a solution found for support problems by backwards induction is a Pareto optimal equilibrium of the associated signaling game. On the basis of this, we will suggest a motivation for the solution concept suggested by Parikh (1992, 2001), who used (a version of) signaling games to account for some conversational implicatures.

The relation between signaling games and support problems is not a one–to–one relation. First, signaling games show more parameters than support problems, especially expectations about the private information of the other player. Hence, there is a whole family of signaling games that can be associated with one support problem. Secondly, a signaling game is associated with many support problems as there are many start nodes in the game, whereas support problems have a fixed start node.

We first define the version of signaling games which is of interest to us and introduce the notion of a perfect Bayesian equilibrium.

Signaling games and perfect Bayesian equilibria

A signaling game is a dynamic game of incomplete information involving two players: a sender (s) and a receiver (r). Formally, a signaling game is a tuple $\langle \{s, r\}, T, p, (F, \mathcal{A}), (U_s, U_r) \rangle$ with the following dynamics (taken from Gibbons, 1992):

1. Nature draws a type t_i for the sender from a set of feasible types $T = \{t_0, \ldots, t_n\}$ according to a probability distribution $p(t_i)$, where $p(t_i) > 0$ for every i and $p(t_0) + \cdots + p(t_n) = 1$.

2. The sender observes t_i and then chooses a message f_j from a set of feasible messages $F = \{f_1, \ldots, f_m\}$.

3. The receiver observes f_j (but not t_i) and then chooses an action a_k from a set of feasible actions $\mathcal{A} = \{a_1, \ldots, a_l\}$.

4. Payoffs are given by $U_s(f_j, a_k, t_i\}$ and $U_r(f_j, a_k, t_i)$, functions from forms, actions, and types to real numbers.

A player's strategy is a complete plan of action: a plan that specifies a feasible action in every contingency in which the player might be called upon to act. A pure strategy for the sender is therefore a function S specifying which message, or form, will be chosen for each type that nature might draw. A pure strategy for the receiver is a function R specifying which action will be chosen for each message/form that the sender might send. For each sender-receiver strategy combination (S_i, R_j) we can determine the *expected* payoffs of the participants given that the sender is of a particular type t — $EU_e(S_i, R_j, t)$ for $e \in \{s, r\}$ — in terms of this participant's utility

functions U_s and U_r. To see this, recall that $S_i(t)$ is a message/form, an element of F, while $R_j(S_i(t))$ is an action, an element of $\mathcal{A}$.

$$EU_e(S_i, R_j, t) = \sum_{t' \in T} \mu_e(t'|S_i(t)) \times U_e(S(t), R_j(S_i(t)), t'), \tag{5.19}$$

where $\mu_e(t'|S_i(t'))$ is defined (in simple signaling games) by means of conditionalization in terms of strategy S_i and (not yet defined) probability distribution P_e that agent e assigns to the selected type as follows (where $S_i^{-1}(f)$ is the set of types in which a sender playing strategy S_i uses message f):

$$\mu_e(t'|S_i(t)) = P_e(t'|S_i^{-1}(S_i(t))). \tag{5.20}$$

A solution of a signaling game is a sender-receiver strategy pair (S, R) and is called a *perfect Bayesian equilibrium*. Such an equilibrium is basically a Nash equilibrium, meaning that both strategies respond optimally, in terms of expected utility, to one another. The optimal strategy for the sender in t is very easy to determine, because he knows of which type he is: $P_s(t) = 1 = \mu_s(t|S_i(t))$ iff s is of type t, 0 otherwise. As a result, his expected utility in t is the same as his actual payoff in $t : EU_s(S_i, R_j, t) = U_s(S_i(t), R_j(S_i(t)), t)$, and thus $S(t)$ must solve $max_{f \in F} U_s(f, R(f), t)$. The optimal strategy for the receiver involves her incomplete information about the sender's type, represented by probability distribution $P_r = p$. Learning that f has been uttered means that the receiver learns that the actual sender's type t is such that $S(t) = f$. Hence, her posterior expectations about the speaker's type result from updating p with $S^{-1}(f)$. Now, R is an optimal strategy for the receiver if her response $R(f)$ to message f chosen by the sender has maximal expected utility given the posterior probability μ, for each f. Thus, the action $R(f)$ must solve $max_{a \in \mathcal{A}} \sum_{t' \in T} \mu_r(t'|f) \times U_r(f, a, t')$. To summarize:

Definition 5.1 *A strategy pair (S, R) is a perfect Bayesian equilibrium for a signaling game $\langle \{s, r\}, T, p, (F, \mathcal{A}), (U_s, U_r) \rangle$ iff:*

1. $\forall t \in T : S(t) \in argmax_{f \in F} U_s(f, R(f), t)$,
2. $\forall f \in F : R(f) \in argmax_{a \in \mathcal{A}} \sum_t \mu(t|f) \times U_r(f, a, t)$,

where $\mu(t|f) = p(t)/p(S^{-1}(f))$, if $S(t) = f$, 0 otherwise.[13]

Representing support problems as signaling games

We now show how to represent support problems as signaling games. It is already obvious that support problems and signaling games have much in common: two participants (E and I, or s and r), with their own action sets (F and $\mathcal{A}$) and their own utility functions (U_E and U_I, or U_s and U_r). What differs is that support problems have probability distributions P_E and P_I, and the semantic interpretation function $[\![\cdot]\!]$, while in signaling games the set of types, T, and the probability distribution p play a crucial role. Recall, first, that each element t of the set of types T represents in a signaling game the sender's, or expert's, private information when the game starts. But this is exactly what is represented by the probability distribution P_E in a support

[13] In this paper we don't care what $\mu(t|f)$ is in case f is not uttered by s in any type.

problem. Thus, we should think of a type as the expert's private information. Because the only knowledge not shared by inquirer and expert in a support problem is the latter's type, we can define the distribution P_E in terms of what the inquirer expects: $P_E(X) = P_I(X|t)$, for any $X \subseteq \Omega$. The semantic interpretation function $[\![\cdot]\!]$ can be straightforwardly modeled in signaling games as a constraint on admissible sender strategies. This leaves us with the parameter p, which represents in signaling games the receiver's/inquirer's expectations about the sender's/expert's type. This parameter is not represented in support problems.[14] Hence, we cannot give a one–to–one mapping from support problems to signaling games. From this it follows that there can also not be a one–to–one mapping between the solutions of support problems and signaling games: to connect the two formalisms we have to prove that the solution found by backwards induction in support problems is independent of p. We will do this by showing that the solution found by backwards induction in support problems give rise to the same behavior as the Pareto optimal perfect Bayesian equilibrium in a signaling game.

For the purpose of comparison we introduce a slightly stricter definition of support problem including some of the previous assumptions needed for calculating optimal answers. We make the tie-breaking rule explicit and we restrict our attention to support problems where both probability measures P_E and P_I are derived from a common prior, i.e. we assume that $P_E(X) = P_I(X|t)$ for $t = \{v \in \Omega \mid P_E(v) > 0\}$. Hence, we can identify the set of all possible speaker's types compatible with P_I with the set $\{t \subseteq \Omega \mid \forall v \in t : P_I(v) > 0\}$.

Definition 5.2 *A* support problem *is a nine–tuple* $\langle \Omega, P_E, P_I, F, (\mathcal{A}, <), U_E, U_I, [\![\cdot]\!] \rangle$ *such that* (Ω, P_E) *and* (Ω, P_I) *are countable probability spaces,* $[\![\cdot]\!]$ *is the semantic interpretation function, and* $\langle (\Omega, P_I), \mathcal{A}, U_I \rangle$ *and* $\langle (\Omega, P_E), F, U_E \rangle$ *are decision problems.* $U_x : F \times \mathcal{A} \times \Omega \longrightarrow \mathbf{R}$ *are the payoff functions (with* U_I *independent of* F*). We assume:*

1. *for* $x = I, E$ *and all* $f \in F$ *and* $a \in \mathcal{A}$: $\sum_{v \in \Omega} P_x(v) \times |U_x(f, a, v)| < \infty$*;*
2. *for all* $X \subseteq \Omega$*:* $P_E(X) = P_I(X|t)$ *for* $t = \{v \in \Omega \mid P_E(v) > 0\}$*;*
3. *(Tie breaking rule)* $<$ *is a linear order on* $\mathcal{A}$*. For* $f \in F$ *we write* $a_f := \max\{a \in \mathcal{A} \mid \forall b \in \mathcal{A} : EU_I(a|[\![f]\!]) \geq EU_I(b|[\![f]\!])\}$*;*
4. $\exists f \in F : EU_E(f) = (\max_{a \in \mathcal{A}} \sum_{v \in \Omega} P_E(v) \times U_I(a, v)) - C(f)$.

Clearly, each support problem in the sense of Definition 5.2 is well-behaved in the sense of Definition 3.1. We denote the set of all support problems as defined now by Σ. As before, we set for a support problem σ: $\mathrm{Op}_\sigma := \{f \in F \mid \forall f' \in F : EU_E(a_f) \geq EU_E(a_{f'})\}$. We use the following notation for support problems σ, σ':

- $\sigma' \sim_I \sigma$ if σ' and σ differ only with respect to P_E.

Given a support problem $\sigma = \langle \Omega, P_E, P_I, F, (\mathcal{A}, <), U_E, U_I, [\![\cdot]\!] \rangle \in \Sigma$, we construct a signaling game $\langle \{s, r\}, T, p, (F, \mathcal{A}), (U_s, U_r) \rangle$ with the following non-trivial identities:

1. $T := \{\sigma' \in \Sigma \mid \sigma' \sim_I \sigma\}$;
2. $U_s(f, a, t) = \sum_{v \in \Omega} P_{E(t)}(v) \times U_{E(t)}(f, a, v)$, with $P_{E(t)}(v) := P_I(v|t)$;

[14] Or better, is represented only in case the expert is known to have complete information of the world he is in.

3. $U_r(f,a,t) = \sum_{v\in\Omega} P_I(v|t) \times U_I(a,v)$;

4. p may be arbitrary as long as $p(\sigma_t) > 0$, where $\sigma_t = \{v \in \Omega | P_{E(t)}(v) > 0\}$ is now thought of as a type.

In order to represent the Gricean maxim of Quality, we have to assume that the following condition holds for the speaker's strategies:

$$\text{For any } t \in T: \text{ if } S(t) = f, \text{ then } t \subseteq [\![f]\!]. \tag{5.21}$$

This condition on strategies for signaling games is equivalent to the previous constraint on support problems stating that the expert can only choose admissible assertions.

For the given signaling game, backwards induction yields all strategy pairs (S,R) with $S(t) \in \mathrm{Op}_{\sigma_t}$ and $R(f) = a_f$. We have to show that (S,R) is a Pareto optimal perfect Bayesian equilibrium.

Proposition 5.3 *Let σ be a given support problem and G an associated signaling game. Then, all strategy pairs (S,R) with*

$$S(\sigma_t) \in \mathrm{Op}_\sigma \text{ and } R(f) = a_f \tag{5.22}$$

are perfect Bayesian equilibria of G. They Pareto dominate all other equilibria.

Proof: We first prove that $\forall t : S(t) \in argmax_{f\in F} U_s(f, R(f), t)$:

$$\begin{aligned} U_s(f,R(f),t) &= \sum_{v\in\Omega} P_{E(t)}(v) \times U_{E(t)}(f,R(f),v) \\ &\leq \max_{f'\in F} \sum_{v\in\Omega} P_{E(t)}(v) \times U_{E(t)}(f',a_{f'},v) \\ &= \sum_{v\in\Omega} P_{E(t)}(v) \times U_{E(t)}(S(t),R(S(t)),v) \\ &= U_s(S(t),R(S(t)),t). \end{aligned}$$

Next we have to show that

$$\forall t_0 \in T : R(S(t_0)) \in argmax_{a\in\mathcal{A}} \sum_t \mu(t|S(t_0)) \times U_r(S(t_0),a,t_0).$$

Let t_0 be given; then:

$$\begin{aligned} \sum_{t\in T} \mu(t|S(t_0)) \times U_r(S(t_0),a,t) &= \sum_{t\in\{t|S(t)=S(t_0)\}} \mu(t|S(t_0)) \times EU_I(a|t) \\ &\leq \max_{a\in\mathcal{A}} \sum_{t\in\{t|S(t)=S(t_0)\}} \mu(t|S(t_0)) \times EU_I(a|t) \\ &= \sum_{t\in\{t|S(t)=S(t_0)\}} \mu(t|S(t_0)) \times EU_I(a_{S(t)}|t) \\ &= \sum_{t\in\{t|S(t)=S(t_0)\}} \mu(t|S(t_0)) \times EU_I(R(S(t))|t). \end{aligned}$$

As $R(S(t)) = R(S(t_0))$ for $t \in \{t \in T | S(t) = S(t_0)\}$, it follows that the last line is identical to

$$\sum_{t \in T} \mu(t|S(t_0)) \times EU_I(R(S(t_0))|t) = \sum_{t \in T} \mu(t|S(t_0)) \times U_r(S(t_0), R(S(t_0)), t).$$

This shows that (S, R) is perfect Bayesian. That it Pareto dominates all other strategy pairs (S', R') follows from

$$\begin{aligned} U_s(S'(t), R'(S'(t)), t) &= \sum_{v \in \Omega} P_{E(t)}(v) \times U_{E(t)}(S'(t), R'(S'(t)), v) \\ &\leq \max_{f \in F} \sum_{v \in \Omega} P_{E(t)}(v) \times U_{E(t)}(f, a_f, v) \\ &= \sum_{v \in \Omega} P_{E(t)}(v) \times U_{E(t)}(S(t), R(S(t)), v) \\ &= U_s(S(t), R(S(t)), t). \end{aligned}$$

and $EU_I(a|t) = \sum_{v \in \Omega} P_E(v) \times U_I(a, v)$. ■

Earlier game theoretical models

In the previous section we gave a game theoretic account of many conversational implicatures, including Horn's division of pragmatic labor. It is exactly implicatures of this latter type that Parikh's (1992, 2001) game theoretic analysis of communication concentrates mostly (if not entirely) his attention on.[15] Using, in essence, standard signaling games as introduced above, together with the assumption that each message f has a fixed semantic meaning, $[\![f]\!]$, Parikh shows that we can account for Horn's division if we choose as the solution concept the Pareto optimal Nash equilibrium. Unfortunately, this Pareto optimal Nash equilibrium is not a standard solution concept in game theory, and so it is not completely clear why it should be used here.

Van Rooij (2004b) suggests that because Horn's division of pragmatic labor involves not only language use but also language organization, one should look at signaling games from an evolutionary point of view. Some natural variants of (standard) evolutionary game theory indeed predict that only the Pareto optimal solution of a cooperative game is evolutionarily stable. Natural as this solution might be (according to at least one of the authors), it cannot explain the implicatures where it is obviously language *use* and online computation that is at issue. Notice that if we assume that the sender is fully informed about the real world (as assumed in sections 4.2 and 4.3), the inverse function S^{-1} of sender strategy S which is part of the Pareto optimal Nash equilibrium (S, R) gives rise to the same pragmatic interpretation as the interpretation rule $Prag_3$ as extensively discussed in the previous section. Based on proposition 5.3, we propose that (at least) for these cases we can motivate the Pareto optimal Nash equilibrium as the most natural solution concept by assuming that the speaker should provide the optimal assertion.

The use of decision and game theoretic ideas to account for conversational implicatures is not new, or limited to the work we mentioned in this paper. First, there is a

[15] In some of the implicatures treated by Parikh (2001) — e.g., the example where the receiver should conclude from 'It is 4 p.m.' that she should go to the talk —, he assumes it is crucial to make use of what he calls 'the value of information'. We would question this assumption, however, and argue that also Parikh himself treats these basically as Q_2-implicatures.

whole battery of papers in biology and economics that make use of signaling games to explain why, or under which circumstances, only true information is communicated. Papers with the same concern, but closer to linguistics are Asher & Williams (1999), and Stalnaker (2006). More closely related to this paper (and other earlier work of the authors) are recent papers by Merin (1999), papers in the edited volume Benz et al. (2006) like Stalnaker (2006), and Jäger (ms) and Van Rooij (to appear). Merin (1999) seeks to explain a number of implicatures on the assumption that speaker and hearer essentially disagree about the desired outcome. Especially the latter two papers are interesting for our present concerns, because they show that we don't have to make use of Pareto optimal Nash equilibria to account for Q_1-implicatures; the standard solution concept of signaling games will do.

6 Conclusion

Psychological evidence suggests that the calculation of conversational implicatures crucially depends on context, and should not be based on a set of mutually given 'linguistic' default rules giving rise to so-called 'generalized implicatures'. In this paper we have taken this suggestion as far as possible: implicatures should all depend on the (perhaps very idiosynchratic) beliefs and common preferences of speaker and hearer, plus the assumption that the speaker is optimally cooperative when he makes his assertion. We have shown how we can implement this view in a uniform way by making use of game theory. We assumed that dialogues are embedded in decision problems, and used backwards induction for calculating optimal assertions. The theory can account for relevance implicatures and mention-some readings of answers, and we have shown how more standard implicatures could be derived by making use of the same uniform theory as well.

In this paper we have made some crucial simplifying assumptions. Most obviously, that the speaker is (i) fully informed about the hearer's decision problem, and (ii) fully cooperative, i.e., shares the hearer's utility function. The first assumption is obviously highly artificial, but we don't think that by giving it up we would radically change our framework. The second assumption might be uncontroversial among proponents of Gricean pragmatics, but it is one that obviously cannot be maintained in general. Fortunately, we have seen that we can think of our theory in terms of signaling games. Signaling games are widely discussed in economics and biology, and our assumption of full cooperation between speaker and hearer is in this framework the exception, rather than the rule. Thus, we can use signaling games to study not fully cooperative communicative situations as well, but in that case we can't assume anymore that the hearer can determine what is implicated by the speaker's utterance by means of our rule of optimal assertion. Thus, we propose that instead of following the Gricean maxims of conversations, the best way to assure that speakers conform to Grice's own cooperativety principle is to assume that speakers act optimally on the assumption that hearer's update their beliefs as described in this paper.

References

[35] Asher, N., I. Sher, and M. Williams (1999), 'Game-theoretical foundations for gricean constraints', in *Proceedings of the Thirteenth Amsterdam Colloquium*, ILLC, Amsterdam, pp. 31-36.

[35] Benz, A. (2006), 'Utility and Relevance of Answers', in A. Benz, G. Jäger, R. van Rooij (eds.), *Game Theory and Pragmatics*, Palgrave, Macmillan, pp. 195-221.

[35] Benz, A., G. Jäger, and R. van Rooij (2006), *Game Theory and Pragmatics*, Palgrave, Macmillan.

[35] Blume, A., Y.G. Kim and J. Sobel (1993), 'Evolutionary stability in games of communication', *Games and Economic Behavior*, **5**: 547-575.

[35] Bott. L. and I. Noveck (2004), 'Some utterances are underinformative: the onset and time course of scalar inferences', *Journal of Memory and Language*, **51**: 437-457.

[35] Breheny, R., N. Katsos, and J. Williams (2006), 'Are generalized scalar implicatures generated by default?', *Cognition*, **100**: 434-463.

[35] Carston, R. (1998), 'Informativeness, relevance and scalar implicature', in R. Carston & S. Uchida (eds.), *Relevance Theory: Applications and Implications*, John Benjamins, Amsterdam, pp. 179-236.

[35] Chierchia, G., S. Crain, M. Guasti, A. Gualmini, & L. Meroni (2001), 'The aquisition of disjunction: Evidence for a grammatical view of scalar implicatures', *Proceedings from the 25th Annual BUCLD*, Sommerville, MA: Cascadilla Press, pp. 157-168.

[35] Clark, R. and M. Grossman (to appear), 'Number sense and quantifier interpretation', *Topoi*.

[35] Fudenberg, D. and J. Tirole, (1991), *Game Theory*, The MIT Press, Cambridge, MA.

[35] Gazdar, G. (1979), *Pragmatics*, Academic Press, London.

[35] Geurts, B. (1998), 'Scalars', in P. Ludewig & B. Geurts (eds.), *Lexikalische Semantik aus kognitiver Sicht*, Gunter Narr, Tübingen, pp. 95-118.

[35] Gibbons, R. (1992), *A Primer in Game Theory*, Harvester Wheatseaf, New York.

[35] Grice, H.P. (1967), 'Logic and conversation', *William James Lectures*, Harvard University, reprinted in *Studies in the Way of Words*, 1989, Harvard University Press, Cambridge, Massachusetts.

[35] Hirschberg, J. (1985), *A theory of scalar implicature*, Ph.D. thesis, UPenn.

[35] Horn, L. (1972), *The semantics of logical operators in English*, Ph.D. thesis, Yale University.

[35] Horn, L. (1984), 'Towards a new taxonomy of pragmatic inference: Q-based and R-based implicature', in Schiffrin, D. (ed.), *Meaning, Form, and Use in Context: Linguistic Applications*, GURT84, Georgetown University Press, Washington, pp. 11-42.

[35] Horn, L. (2004), 'Implicature', in L. Horn & G. Ward (eds.), *Handbook of Pragmatics*, Oxford: Blackwell, pp. 3-28.

[35] Jäger, G. (2006), *Game dynamics connects semantics and pragmatics*, ms. University of Bielefeld.

[35] Kempson, R. (1986), 'Ambiguity and the semantics-pragmatics distinction', in C. Travis (ed.), *Meaning and Interpretation*, Blackwell, Oxford, pp. 77-103.

[35] Levinson, S. *Presumptive Meanings. The Theory of Generalized Conversational Implicatures*, MIT Press, Cambridge, Mass.

[35] Lewis, D. (1969), *Convention*, Cambridge: Harvard University Press.

[35] Merin, A. (1999), 'Information, relevance, and social decision making: Some principles and results of decision-theoretic semantics, in L. Moss et al. (eds.) *Logic, Language, and Information*, volume 2, CSLI Publications, pp. 179-221.

[35] Noveck, I. (2001), 'When children are more logical than adults: experimental investigations of scalar implicatures', *Cognition*, **78**: 165-188.

[35] Noveck, I. and A. Posada (2003), 'Characterizing the time course of an implicature: An evoked potential study', *Brain and Language*, **85**: 203-210.

[35] Papafragou, A. and J. Mussolino (2003), 'Scalar implicatures: experiments at the semantic/pragmatics interface', *Cognition*, **86**: 253-282.

[35] Parikh, P. (1992), 'A game-theoretical account of implicature', in Y. Vardi (ed.), *Theoretical Aspects of Rationality and Knowledge: TARK IV*, Monterey, California.

[35] Parikh, P. (2001), *The use of Language*, CSLI Publications, Stanford, California.

[35] Storto, G. and M.K. Tannenhaus (2004), 'Are scalar implicatures computed online?', in *Proceedings of WECOL 2004.*

[35] Rooij, R. van (2003), 'Questioning to Resolve Decision Problems', *Linguistics and Philosophy* **26**: 727–763.

[35] R. van Rooij (2004a), 'Relevance and Bidirectional Optimality Theory', in R. Blutner and H. Zeevat (eds.), *Optimality Theory and Pragmatics*, Palgrave MacMillan, Hampshire, pp. 173-210.

[35] Rooij, R. van (2004b), 'Signalling games select Horn strategies', *Linguistics and Philosophy*, **27**: 493-527.

[35] Rooij, R. van (to appear), 'Optimality Theoretic and Game Theoretic Approaches to Implicatures', *Stanford Encyclopedia of Philosophy.*

[35] Rooij, R. van, and K. Schulz (2004), 'Exhaustive interpretation of complex sentences', *Journal of Logic, Language and Information*, **13**: 491–519.

[35] Stalnaker, R. (2006), 'Saying and meaning, cheap talk and credibility', in A. Benz, G. Jäger, R. van Rooij (eds.), *Game Theory and Pragmatics*, Palgrave, Macmillan, pp. 83-100.

Anton Benz
University of Southern Denmark, Kolding
Engstien 1
DK 6000 Kolding
ab@anton-benz.de

Robert van Rooij
Institute for Logic, Language and Computation
Universiteit van Amsterdam
Nieuwe Doelenstraat 15
1012 CP Amsterdam
r.a.m.vanrooij@uva.nl