Quality and Quantity of Information Exchange

Robert van Rooy

Abstract. The paper deals with credible and relevant information flow in dialogs: How useful is it for a receiver to get some information, how useful is it for a sender to give this information, and how much credible information can we expect to flow between sender and receiver? What is the relation between semantics and pragmatics? These Gricean questions will be addressed from a decision and game-theoretical point of view.

Keywords: Gricean pragmatics, relevance, meaning, signaling games

1. Introduction

Within linguistic *pragmatics*, Grice's (1967) *cooperative principle* has always played an important role: the assumption that speakers are maximally efficient, rational cooperative language users. Grice lists four rules of thumb – the maxims of *quality*, *quantity*, *relevance*, and *manner* – that specify what participants have to do in order to satisfy this principle. They should speak sincerely, relevantly, and clearly and should provide sufficient information. Notwithstanding its merits and influence, Gricean pragmatics leaves something to be desired. The most obvious shortcoming of the theory is that its key notions are stated in a despairingly vague way. Grice's formulation of the *relevance* maxim, for instance, Be Relevant!, is somewhat disappointing. More disturbing, perhaps, is Grice's idealization of the conversational agents as fully cooperative. Agents each come to a conversation with their own goals and preferences and these coincide only in special cases. Although language seems to be a multi-purpose instrument, a basic purpose is to influence other's behavior in accordance with one's own preferences. This suggests that *strategic* considerations play a much more important role in language use than recognized by Grice. If language is used to influence others, this persuasion might involve lying and/or giving *misleading* information. This suggests that the traditional Gricean issues concerning the *quantity*, *quality*, and *relevance* of (expected) information exchange should be considered from a broader perspective. Taking such a broader perspective has consequences for the relation between the *conventional* meaning of an expression and its *conversational* impact: what the speaker actually means by her use of an expression depends on her goals and preferences.

In this paper I will take such a broader perspective on language use and address the above mentioned issues from a decision and game-theoretical point of view.[1] The paper deals with credible and relevant information flow in dialogs: How useful is it for a receiver to get some information, how useful is it for a sender to give this information, and how much credible information can we expect to flow between sender and receiver? I will also investigate the relation between the semantics and pragmatics of natural language: does semantics have primacy over pragmatics or is it rather the other way around? The purpose of this paper is not so much to state new results (although there are some, particularly in sections 3 and 6), but to show that insights of one discipline (economics) can help to resolve long-standing issues in another discipline (linguistics), or at least that these old questions can be looked upon from a new, broader, and therefore (hopefully) refreshing point of view.

In the first part of this paper, sections 2 through 4, I will discuss the usefulness of asking questions and receiving information. Thus, this part concerns Grice's maxim of *relevance*. The rest of the paper discusses the amount of possible *credible* information transmission and mainly concerns Grice's maxim of *quality*. Game theory is used here to account for the strategic reasoning in communicative behavior. Grice's maxim of *quantity* plays an important role throughout, as does the relation between the semantics and pragmatics of natural language.

2. From informativity to relevance: questions and assertions

To know the meaning of a sentence, you have to know under which circumstances this sentence is true. This naturally gives rise to the stronger proposal to *equate* the meaning, or content, of a sentence with its truth conditions, represented by the set of situations in which it is true, i.e. the proposition expressed by that sentence. While such an analysis might be natural for *declaratives*, sentences used to make *assertions*, it is not so for *interrogatives*, sentences that express *questions*. Still, the latter can be analyzed along similar lines. According to one approach, you know the meaning, or content, of a question when you know what counts as a resolving answer. On the assumption that the possible resolving answers are complete and therefore mutually exclusive, it follows that a question should be represented by a *partition*. Groenendijk & Stokhof (1984) argue for a partitional view of questions on the basis of linguistic phenomena, but partitions

[1] See also P. Parikh (1991, 2001), R. Parikh (1994), and Merin (1999) for taking a similar perspective. These were all an important impetus for my own work.

have always played an important role within statistical decision theory (Savage, 1954) to determine the value of an experiment, a question to nature. If questions are represented by partitions, they can be compared naturally as follows: question Q is *better* than question Q' iff for every element of Q, there must be an element of Q' such that the former entails the latter, i.e. $Q \sqsubseteq Q'$

$$Q \sqsubseteq Q' \quad \text{iff} \quad \forall q \in Q : \exists q' \in Q' : q \subseteq q'$$

In fact, this is Groenendijk & Stokhof's (1984) entailment relation between questions.

We may determine how 'good' questions and assertions are in terms of entailment relations. In fact, this is the standard way of doing so within linguistic pragmatics. Grice's maxim of quantity, for instance, which asks the speaker to be as informative as possible, is normally formalized in terms of entailment. However, the entailment relations are sometimes too fine-grained, and at other times, not fine-grained enough for this aim. Entailment is too fine-grained, because it is clear that speakers should not be more informative than *required* for the purpose of the conversation (cf. Grice's second submaxim of quantity). Thus, we require a more context-dependent notion of 'relevance'. The relations are not fine-grained enough, because we have more intuitions about 'relevance' than these entailment relations can capture. Sometimes a question or an answer can, intuitively, be more relevant than another, although the former does not entail the latter. What this suggests is that (i) questions and assertions should be compared to each other with respect to a more quantitative ordering relation, but also that (ii) to compare the usefulness of two questions or two assertions with each other, we should *relate* these to (something like) a *third question*, or a *decision problem.*

3. Decision Criteria and the Value of Information

The idea is that we ask questions not just to get new information, but rather to get new information that might help to resolve a particular decision problem. By relating questions to decision problems, we can measure the utility of questions and answers/assertions. It turns out that this measure depends crucially on the *decision criterion* used.

Decision problems are conventionally categorized according to the decision maker's knowledge of the state of nature (cf. Luce & Raiffa, 1957) and her preferences. A problem is *under strict uncertainty* when the agent's knowledge state is consistent with a number of situations, and she cannot quantify her uncertainty about which of those situations is the real one. In decisions under strict uncertainty and *ordinal* preferences, the decision problem can be modeled by a triple like $\langle T, E, \geq \rangle$, where (i) T is the set of states that the agent thinks are possible, representing her beliefs, (ii) E is the set of alternative actions that she considers, and (iii) $\geq$ is a partial order on state-action pairs, i.e. $\geq \subseteq T \times E$, which represents her preferences. Following Savage (1954), I will take the actions in set E to be primitives. When our agent can represent her preferences by a *cardinal* utility function, the ordering $\geq$ is replaced by U, a utility function from state-action pairs to real numbers. When the agent *can* quantify her uncertainty, the problem is standardly called a *decision under risk*. In these cases a decision problem contains an additional probability function P which assigns to states their (subjective) probabilities.

These different ways of representing decision problems gave rise to different ways to resolve them. The pro's and con's of these decision criteria have been discussed by several authors in terms of certain postulates of rationality.[2] I won't go into these discussions, but will only define the usefulness of assertions and questions in terms of them, and evaluate their predictions. In this I will follow Savage (1954) in discussing what he calls the *value of observations*. It turns out that the different decision criteria don't agree on whether the value of information of assertions behave monotone increasingly with respect to entailment, or under which circumstances questions are predicted to be useless.

One of the goals of this paper is to become clear about the relation between the semantics and pragmatics of natural language: does semantics have primacy over pragmatics or is it rather the other way around? In the next (sub)sections I will discuss to what extent semantic *entailment* relations between propositions and questions can be seen as *abstractions* from corresponding *pragmatic* utility-based relations between them.

3.1. Decision with strict uncertainty: ordinal preferences

Suppose an agent's decision problem is represented by a triple like $\langle T, E, \geq \rangle$ as described above. Which action should our agent choose? This is clear when there is one *dominating* action: an action that is preferred to all others in all situations. The existence of such a dominating action is rare, however. Still, also in case there is no such an action we can say what our agent should *not* do: she should not perform an action that is (strictly) *dominated* by another action. Let's define $\mathcal{O}^o(A)$, the set of potentially *optimal* actions, as the set of non-dominated actions

[2] See Luce & Raiffa (1957), for instance.

after A is learned: $\{e \in E | \neg\exists e' \in E : \forall t \in A : \langle t, e'\rangle > \langle t, e\rangle\}$ (where '$>$' is defined as usual in terms of '$\geq$'). Notice that this set may be smaller, but certainly not larger, than the set of actions that might be optimal after you learn the trivial proposition, i.e. $\mathcal{O}^o(\top)$. We say that proposition A gives *relevant* information just in case the set of non-dominated actions after learning A strictly decreases. Similarly, we say that A is *more relevant* than B with respect to decision problem DP, $A >^o_{DP} B$, iff $\mathcal{O}^o(A) \subset \mathcal{O}^o(B)$, i.e., learning A helps more to resolve the decision problem than learning B does.

The set of alternative actions, E, gives rise to a set of propositions. We can relate each action $e \in E$ to the set of situations in which there is no other action e' in E that is strictly better. We will denote the proposition corresponding with e by e^* and the resulting set of propositions by E^*. The set of propositions E^* can be thought of as a question (What should I do?), although it does not in general partition the state space. Now the above induced ordering relation comes down to the claim that A is better to learn than proposition B just in case the set of actions that are potentially optimal after learning A is a subset of the potentially optimal actions after learning B:

$$\mathcal{O}^o(A) \subset \mathcal{O}^o(B) \quad \text{iff} \quad \{e^* \in E^* : \; A \cap e^* \neq \emptyset\} \subset \{e^* \in E^* : \; B \cap e^* \neq \emptyset\}$$

It is worth remarking that, in this way, we have reduced the ordering of propositions in terms of decision problems to the ordering between answers that Groenendijk & Stokhof (1984) have proposed and applied to account for some linguistic phenomena.

In general, it obviously holds that if $A \subseteq B$, then also $\mathcal{O}^o(A) \subseteq \mathcal{O}^o(B)$. Thus, new information can never be undesirable in the sense that it increases the uncertainty of the decision (i.e., makes more actions potentially optimal). Although the reverse does not hold, something more general is the case: if $\mathcal{O}^o(A) \subseteq \mathcal{O}^o(B)$ holds for *every* decision problem, it also is the case that $A \subseteq B$. Thus, we have the following

Fact: $\quad A \subseteq B \quad \text{iff} \quad \forall DP : A \geq^o_{DP} B$.[3]

Can we now also define a comparative ordering relation between *questions*? The following proposal is straightforward: $Q \geq^o Q'$ iff $\forall q \in Q : \exists q' \in Q' : q \geq^o q'$. From the fact proved above and the definition of

[3] **Proof**: We know already that if $A \subseteq B$ then for all decision problems DP: $A \geq^0_{DP} B$. To prove the other way round, suppose that $\forall DP : A \geq^o_{DP} B$, but $A \not\subseteq B$. Then $\exists t \in A : t \notin B$. But then we can think of a relation $>$ and an action e such that $\forall e' \neq e : \langle t, e\rangle > \langle t, e'\rangle$ and $\forall t' \in B : \langle t', e'\rangle > \langle t', e\rangle$. Thus, $e \in \mathcal{O}^o_{DP}(A)$, but $e \notin \mathcal{O}^o_{DP}(B)$, which is in contradiction with what we assumed.

$Q \sqsubseteq Q'$, we immediately have the following

Fact: $Q \sqsubseteq Q'$ iff $\forall DP : Q \geq^o_{DP} Q'$.

The above facts are obviously interesting for the semantic-pragmatic interface: one proposition/question entails another just in case it is *always* more (or equally) useful.

Although the definition of *relevance* given here is quite appealing, we have more intuitions about it than this *qualitative* notion can capture. If my choice between action e_1 and e_2 depends on whether both John and Mary are sick, your assertion *At least John is sick* is compatible with both e_1^* and e_2^*, but still felt to be very relevant. Similarly for questions: If I want to find out who of John, Mary, and Sue are sick, the question *Who of John and Mary are sick?* is felt to be more informative, or relevant, than the question *Is Sue sick?*, although the relevance relation discussed in this section does not compare both questions. In the following sections we will overcome these problems by representing decision problems in a more *quantitative* way.

3.2. Decision under strict uncertainty: Maximin

If a decision maker's preferences between state-action pairs can be represented by a cardinal utility function U, we can make stronger recommendations about which action the decision maker should perform. According to Wald's (1950) *maximin* criterion of solving a decision problem, we should choose that action e which maximizes e's security level. The *security level* of e, $S(e)$, is the utility of action e under the worst possible consequences that the agent can imagine, i.e. that the worst possible state in T is true:

$$S(e) \quad = \quad min_{t \in T} U(t, e)$$

The action to be chosen, i.e. the action with maximal security level, is then $max_{e \in E} S(e)$. Thus, Wald suggests the following decision rule:

choose e_i such that $S(e_i) \quad = \quad max_{e \in E} S(e) \;=\; max_{e \in E} min_{t \in T} U(t, e)$

To illustrate, look at the following decision problem:

U-table	e_1	e_2	e_3
t_1	5	2	3
t_2	6	0	2
t_3	0	3	4

Notice that this matrix has a maximin value of 2, i.e. $max\{0, 0, 2\} = 2 = U(t_2, e_3)$. This decision criterion corresponds with one of an agent choosing first in a zero-sum 2-stage sequential game with complete information. And indeed, the rule is a very pessimistic criterion of choice; its general philosophy is to assume that the worst will happen.

On the assumption that, when confronted with a decision problem, the agent will make the maximin decision, the *decision problem* also has a *value*: the security level of the action which has the maximal security level when the agent has to *choose now*:

$$MV(\text{Choose now}) \quad = \quad max_{e \in E} S(e)$$

In our case $MV(\text{Choose now}) = 2$, since $max_{e \in E} S(e) = U(t_2, e_3) = 2$.

But now suppose that the agent doesn't have to choose now, but has the opportunity to first receive some useful information. Suppose that another participant of the dialogue truthfully asserts A, and that the agent herself updates her belief state with A. Let us now define the security level of e after A is learned, $S(e/A)$, as the utility of e given the worst outcome possible when A is true, i.e. $S(e/A) = min_{t \in (T \cap A)} U(t, e)$. Then we can define the (maximin) value of choosing after A is learned, $MV(\text{Learn } A, \text{ choose later})$, as $max_{e \in E} S(e/A)$:

$$MV(\text{Learn } A, \text{ choose later}) \;=\; max_{e \in E} S(e/A) \;=\; max_e min_{t \in A} U(t, e)$$

If $A = \{t_1, t_2\}$, for instance, this value is $max_{e \in E} min_{t \in \{t_1, t_2\}} \{U(t, e)\} = max\{5, 0, 2\} = U(t_1, e_1) = 5$. Using this value, we can now determine the maximin value of the information A itself as the difference between the value of choosing after learning A, $MV(\text{Learn } A, \text{ choose later})$, and the value of choosing now, $MV(\text{Choose now})$:

$$\begin{array}{rcl} MV(A) & = & MV(\text{Learn } A, \text{ choose later}) - MV(\text{Choose now}) \\ & = & max_{e \in E} S(e/A) - max_{e \in E} S(e) \end{array}$$

For our example where $A = \{t_1, t_2\}$, this value is positive: $MV(A) = 5 - 2 = 3$. We say that receiving proposition A is better than receiving B, $A > B$, iff $MV(A) > MV(B)$. It may appear that the exact numbers are crucial for this ordering relation. However, this is not really the case. The preference ranking on propositions determines the maximin value up to a linear transformation with positive coefficient.[4]

How is this ordering relation related to the one induced by entailment? We will show that the former is at least as fine-grained as the latter. It is easy to see that when $A \subseteq B$, the value of choosing after

[4] Based on the fact that the utility function itself is based on a preference ranking only up to such a linear transformation.

you learn A, MV(Learn A, choose later), is always at least as high as the value of choosing after you learn B, MV(Learn B, choose later). To show this, notice that $\forall e \forall A : min_{t \in A} U(t,e) \geq min_{t \in T} U(t,e)$. From this it follows that $\forall A : max_e min_{t \in A} U(t,e) \geq max_e min_t U(t,e)$. But this means that the value of A is always greater than or equal to 0, $MV(A) \geq 0$, and that if $A \subseteq B$, then $MV(A) \geq MV(B)$. Thus, the value of information according to the maximin decision rule behaves monotone increasingly with respect to the subset relation, '$\subseteq$', between propositions.

Does the inverse of this relation also hold? Of course it won't be the case that whenever $MV(A) \geq MV(B)$ it also holds that $A \subseteq B$. Something more general, however, does hold: If $MV(A) \geq MV(B)$ with respect to *every* decision problem, i.e. with respect to every utility function, it also is the case that $A \subseteq B$.

Fact: $A \subseteq B$ iff $\forall DP : MV_{DP}(A) \geq MV_{DP}(B)$.[5]

Notice that this fact means that in zero-sum games it never does any harm to get information. For the player who has to choose which state is true, this means that in such games it is never useful to give (truthful) information. Indeed, giving truthful information would, in a game-theoretical setting, mean that before the other player makes his move, you would already truthfully *publically commit* yourself to a particular choice of action, which is never a useful thing to do in zero-sum games.

Just as in the previous section, we might also define the usefulness of new information in terms of the set of actions it excludes as potentially optimal. In order to do so let us define $\mathcal{O}^{MV}(A)$ to be $\{e \in E | \exists B \subseteq A : S(e/B) = max_e S(e/B)\}$. According to this usefulness criterion, new information behaves monotone increasing. The set of *possible* optimal actions can only decrease when one receives more information. The stronger **fact** can be proved just as in the previous section: $A \subseteq B$ just in case for every decision problem DP, $\mathcal{O}^{MV}_{DP}(A) \subseteq \mathcal{O}^{MV}_{DP}(B)$.

In accordance with the negative philosophy behind the maximin decision criterion, we may now also define the value of a *question*, $MV(Q)$, as the value of the answer which has the lowest value:

[5] **Proof**: We know already that if $A \subseteq B$ then for all DP, $MV_{DP}(A) \geq MV_{DP}(B)$. Suppose that the other way round is not the case, i.e., that $\forall DP : MV_{DP}(A) \geq MV_{DP}(B)$ but $A \not\subseteq B$. Then $\exists t \in A : t \notin B$. But then we can think of a U such that $\forall e : \forall t' : t' \neq t \rightarrow U(t', e,) > U(t,e)$. This means that $\exists U : max_e min_{t'} U(t',e) = max_e U(t,e)$ and thus that $MV(A) = max_e U(t,e)$. But if $t \notin B$ then $MV(B) > max_e U(t,e) = MV(A)$. This is in contradiction with what we have supposed before. Thus if for all DP, $MV_{DP}(A) \geq MV_{DP}(B)$, then it has to be the case that $A \subseteq B$.

$$MV(Q) \quad = \quad min_{q \in Q} MV(q)$$

Thus, a question is as good as its least useful answer. What is the value of a question like $\{\{t_1, t_3\}, \{t_2\}\}$? In order to calculate this, we first determine for both answers their maximin values: $max_{e \in E} min_{t \in \{t_1, t_3\}} U(t, e) = max\{0, 2, 3\} = U(t_1, e_3) = 3$, and $max_{e \in E} min_{t \in \{t_2\}} U(t, e) = max_{e \in E} U(t_2, e) = U(t_2, e_1) = 6$. If the value of the question is equated with the difference between the value of the answer with the minimal maximin value and the maximin value of the original decision problem, the value of the question, $MV(\{\{t_1, t_3\}, \{t_2\}\})$, is 3 - 2 = 1.

In section 2 we said that question Q is more *fine-grained* than question Q' iff $Q \sqsubseteq Q'$. From the fact that the values of assertions behave monotone increasingly with respect to the subset relation '$\subseteq$', it follows immediately that the values of questions also behave monotone increasingly with respect to the relation '$\sqsubseteq$'. This has the consequence that the most fine-grained question will also have the highest value. However, just as for propositions, we can prove that something more general holds: $Q \sqsubseteq Q'$ iff the maximin value of Q is higher than that of Q' for *every* decision problem.

Fact: $\quad Q \sqsubseteq Q' \quad$ iff $\quad \forall DP : MV_{DP}(Q) \geq MV_{DP}(Q')$.[6]

Now suppose that the question is maximally fine-grained, i.e. that each answer singles out a unique state. In that case, the expected value of the question, i.e. the expected value of *perfect* information is the difference between $min_t max_{e \in E} S(e, t)$ and $max_{e \in E} S(e)$. But this is the same as the difference between the so-called *minimax*-value of a decision problem, i.e. $min_t max_e U(t, e)$, and its *maximin*-value, $max_e min_t U(t, e)$. Because this difference can never be negative, the most fine-grained question will indeed never have a negative value. To illustrate the difference between the minimax- and the maximin-value, look at our matrix representing the decision problem again:

[6] **Proof** (thanks to Katrin Schulz): The proof from left to right follows from what I said earlier. From right to left we prove by contraposition. Suppose $\forall DP : MV_{DP}(Q) \geq MV_{DP}(Q')$ but $Q \not\sqsubseteq Q'$, i.e. $\exists q \in Q, r_1, r_2 \in Q' : q \cap r_1 \neq \emptyset$ and $q \cap r_2 \neq \emptyset$. Now take a DP with three actions: e_1, e_2 and e_3 and U such that $\forall i : min_{t \in T} U(t, e_i)$ lays in q (*). Moreover, suppose that $min_{t \in (q \cap r_1)} U(t, e_1) < min_{t \in (q \cap r_1)} U(t, e_2) < min_{t \in (q \cap r_1)} U(t, e_3)$ and $min_{t \in (q \cap r_2)} U(t, e_1) > min_{t \in (q \cap r_2)} U(t, e_2) > min_{t \in (q \cap r_2)} U(t, e_3)$. Then it follows that $max_{e \in E} min_{t \in r_1} U(t, e) > max_{e \in E} min_{t \in q} U(t, e) < max_{e \in E} min_{t \in r_2} U(t, e)$. Together with assumption (*) we can conclude that $MV_{DP}(Q) < MV_{DP}(Q')$, which is in contradiction with our supposition. This is enough to prove the fact.

U-table	e_1	e_2	e_3
t_1	5	2	3
t_2	6	0	2
t_3	0	3	4

Notice that this matrix has a minimax value of 4, i.e. $min\{5,6,4\} = 4 = U(t_3, e_3)$, and a maximin value of 2, i.e. $max\{0,0,2\} = 2 = U(t_2, e_3)$. This means that the maximin value of the most fine-grained question, $MV(\{\{t_1\},\{t_2\},\{t_3\}\})$, is 4 - 2 = 2. This, of course, is the maximal value a question can have, given this decision problem.

Given a matrix representation of a decision problem as above, $U(t, e)$ is the entry in row t and column e. Adopting some terminology from game-theory, we call the pair $\langle t, e\rangle$ a *saddle point* for the matrix, when $U(t, e)$ is the largest entry in its row, and the smallest entry in its column. Notice that the above decision problem doesn't have a saddle point. What is the value of a question when the original decision problem *does* have a saddle point? A well known game-theoretical fact tells us that when a game has a saddle point, (the *Nash equilibrium* of a zero-sum game), the minimax value *coincides* with the maximin value. But this means that in such cases even the finest-grained question has a value of 0, i.e. the question is predicted to be *useless*. In these cases coarser-grained questions will also have a value of 0. Thus, the maximin-criterion for choosing predicts that asking questions is useless when the 'game' has a saddle point. But this prediction becomes problematic, and very counterintuitive, when we allow for *mixed* actions, since when all mixed-actions are allowed, saddle-points always exist in two-person zero-sum games. This means that questions are *always* predicted to be useless when answers can be mixed acts.[7]

3.3. Decision under risk: expected utilities

In the previous section I assumed that the questioner should always expect the worst response. The result was that too many questions were predicted to be useless. I now claim that the assumption is not natural

[7] Similarly as for the maximin criterion, we can also determine the value of assertions and questions with respect to other decision criteria under strict uncertainty, in particular Savage's (1954) *minimax loss* criterion (cf. Szaniawski (1967) for an analysis of (among others) the minimax loss value of *perfect* information). One can show that the **facts** proved here for maximin decision rule are true for this criterion as well. The consequence of allowing for mixed actions is not as problematic, however, for the minimax loss criterion.

and propose dropping it. Since the respondent gives a *true* answer to the question, why not determine the utility on the basis of the *expectation* one has about which answer is true? Indeed, it seems more natural to assume that the value of the question is the *expected* value of the answer, where expectation is measured in terms of probability.

When the agent can quantify her uncertainty, and we can represent her uncertainty by a *probability* function, the problem is standardly called a *decision under risk*. In these cases a decision problem is modeled as a quadruple like $\langle T, P, E, U\rangle$, where P is a probability function over T. Now we say that the *expected utility* of action e, $EU(e)$, with respect to probability function P is

$$EU(e) \quad = \quad \sum_t P(t) \times U(t,e)$$

If our agent faces a *decision problem* and she has to choose now, she obviously should choose the act with the highest expected utility.

$$UV(\text{Choose now}) \quad = \quad max_{e \in E} \sum_t P(t) \times U(t,e)$$

The utility value of making an informed decision after learning A, UV(Learn A, choose later), is the expected utility conditional on A of the action that has highest expected utility (thus, $max_e EU(e/A)$):

$$UV(\text{Learn } A, \text{ choos later}) \quad = \quad max_e \sum_t P(t/A) \times U(t,e)$$

Now we determine the value, or *relevance*, of the assertion A as follows:

$$UV(A) \quad = \quad UV(\text{Learn } A, \text{ choose later}) - UV(\text{Choose now})$$

One may claim that in a cooperative dialogue one assertion, A, is 'better' than another, B, just in case the utility value of the former is higher than the utility value of the latter, $UV(A) > UV(B)$. However, according to this measure new information might have a negative value: although $A \subseteq B$, it might be that $UV(A) < UV(B)$.

The notion $UV(\cdot)$ is very useful to determine the utility values of *questions*. We say that the *expected* utility value of question Q, $EUV(Q)$, is the *average* utility value of its possible answers:

$$EUV(Q) \quad = \quad \sum_{q \in Q} P(q) \times UV(q)$$

Somewhat surprisingly, perhaps, given the fact that $UV(q)$ can be negative, $EUV(Q)$ cannot be negative. This can be shown using the fact that our notion of $EUV(Q)$ is provably equivalent to the *expected value of sample information*,[8] $EVSI$, which plays an important role in statistical decision theory (cf. Raiffa & Schlaifer, 1961). In fact, the value will only be 0 in case no answer to the question results in the agent changing her mind about which action to perform. In these circumstances the question really does seem irrelevant, and it thus seems natural to say that question Q is *relevant* just in case $EUV(Q) > 0$. It should also be obvious that this measure function totally orders all questions with respect to their expected utility value. As a special case of Blackwell's (1953) theorem, we can prove that if $EUV_{DP}(Q)$ is the expected utility of question Q with respect to decision problem DP, it holds that $Q \sqsubseteq Q'$ iff $\forall DP: \ EUV_{DP}(Q) \geq EUV_{DP}(Q')$. Thus, we can think of the *semantic* entailment relation between questions as an abstraction from the corresponding *pragmatic* utility based relation.[9]

What about the more cautious utility measure that can be defined similarly to the ones we defined in the previous (sub)sections? Let $\mathcal{O}^{EU}(A)$ be the set of possible optimal actions after one learns that A is the case: $\{e \in E | \exists B \subseteq A : EU(e/B) = max_e EU(e/B)\}$. As one might expect, this utility measure is monotone with respect to informativity.

3.4. A brief comparison

In this section I have discussed some utility measures of assertions and questions with respect to various types of decision situations. The difference between, on the one hand, the purely qualitative decision situation, and the ones where we have more quantitative information, on the other, is that in the latter case a more definite recommondation can be given. However, this recommendation is risky, even when the beliefs are not represented by a probability function. Define $\mathcal{OPT}(A)$ to be the set of actions optimal if A is the case, i.e. $\mathcal{OPT}(A) = \{e' \in E | V(e'/A) = max_e V(e/A)\}$, where V is either 'S' or 'EU'. Then neither is monotone increasing in the following sense: $A \subseteq B \rightarrow \mathcal{OPT}(A) \subseteq \mathcal{OPT}(B)$. However, we saw that with respect to $\mathcal{O}^V(A)$ (V is either 'MV' or 'EU'), both decision rules behave monotone: the set of *possible* optimal actions can only decrease when one receives more information. In this sense, the value of information behaves the same in the two quantitative decision situations as in the qualitative one. We also saw that, when we abstract away from the particular decision situation, an

8 See van Rooy (to appear).

9 For a discussion of the relation between, on the one hand, the *utility* values of assertions and questions as discussed in this section and, on the other hand, some *information* values definable in terms of Shannon's (1948) communication theory, see e.g. van Rooy (to appear).

ordering in terms of the measure $\mathcal{O}$ reduces to (or characterizes) the context-independent entailment relation in all cases considered.

In those cases where a definite recommodation could be made, we could also determine the utility value of (expected) new information. However, we saw an important difference between the maximin and expected utility decision rules with respect to the utility of *assertions*: in contrast to maximin, deciding by expected utility has the effect that new information can, from a first-person point of view, decrease the (expected) utility of the (at that moment) preferred action. This is due to the fact that deciding by expected utility is, in a sense, less *cautious* than deciding by maximin. The extra risk involved if one takes probabilities seriously is that due to the high probability of favorable situations, learning that a less favorable situation is true can have more damaging effects.

In this section we saw that although the maximin decision criterion for assertions doesn't seem unnatural, the use of the criterion to measure the value of *questions* is based on the counterintuive assumption that the answerer will give the worst possible response. The unnaturalness of the measure is reflected by the (many) circumstances in which questions are predicted to be useless. This is why we introduced the notion that measures the expected utility of a questions, which worked much more in accordance with intuition. However, the following section will show that an overly simplistic use of the expected utility measure leads to counterintuitive predictions concerning the value of questions.

4. Information hurts

Intuitively, an informed decision maker is better off than one with less information. And, indeed, for the decision rules discussed in the previous section, we have seen that one can always expect better results from making decisions on more knowledge rather than less. Receiving information was modeled by eliminating states that represent what one believes. A state, or world, represents everything that is relevant to the decision maker to determine her action. In particular, it should represent the actions that other agents perform that are relevant for her payoff. The actions of the other agents depend, of course, partly on their beliefs. But this means that, in our way of determining the (expected) utility of new information, we have made an important assumption: if our decision maker learns something, other agents that can make decisions that are payoff-relevant for the actions of our agent herself do not learn anything. As we will see in this section, if we drop this assumption, getting more information can be harmful.

4.1. Information rejection in games

In the previous section we have shown that it is always considered to be useful for an agent to ask for information. Paradoxically enough, however, in *games* this doesn't seem to be the case. In some game-like situations one or all participants of a game might *refuse* to get more information, on the grounds that it would hurt them.[10] In this section we will informally discuss some such examples borrowed from Hirshleifer & Riley (1992) and see how to resolve this paradox.

Rejection of *public* information. Suppose three identical individuals are offered the following deal. Provided there is unanimous agreement to the proposal, two of them will receive a payoff of three euros, but one of them will get nothing. If any one of them doesn't agree, they get one euro each. If it is commonly assumed that the chance for being the unlucky person is $\frac{1}{3}$, expected utility theory predicts that they should agree to the proposal. But, before that decision is actually effectuated, the persons are offered another option: they can choose to be informed in advance, without charge, as to who the unlucky person would be. Evidently, if the three were unanimously willing to accept the initial gamble, they must now be unanimously *unwilling* to accept the information even if given for free – since receiving it would destroy the possibility of unanimous agreement on the proposal.

Rejection of *private* information. Two individuals i and j are to predict whether the top card of a shuffled deck is black or red. The predictions are made in sequence: i guesses first, and then j (after hearing i's guess). If they make the same prediction, each wins 1 euro, whether their prediction turns out correct or not. If they disagree, the one who is correct wins 3 euros, and the other nothing. Since i has no basis for preferring one or the other color, he will choose at random. Then j will choose the opposite color, so each will have an expected gain of 1.50 euros. Now suppose someone offers, without charge, an arrangement whereby the first player i is told the actual color of the top card before he makes his prediction. Evidently, the parties would be unanimous in rejecting that arrangement – even i as the "beneficiary." For, with that information, individual i would choose the correct color, j would then make the same choice, and they would each gain only 1 euro rather than an expectation of 1.50 euros.

As explained by Hirshleifer & Riley (1992), in both examples, the information would be rejected because, while there is a *latent* conflict of interest among the players, that conflict is not relevant for the decisions made in a state of ignorance. Disclosure of the information,

[10] See, for example, exercise 28.2 of Osborne & Rubinstein (1994).

however, makes the latent conflict of interest an *actual* one, which is disadvantageous for all players involved.

The fact that information may hurt the agent who receives it may be counter-intuitive at first sight and in conflict with Blackwell's theorem mentioned in section 3.3. However, this conflict is not real. As explained by Neyman (1991), more information is always valuable to an agent, if everything else remains the same, in particular, if *the others are not aware of the fact that the agent receives new information*. Otherwise, the other agents may behave in a way that hurts the informed one.[11]

Notice that in both examples above all involved players learn something. In the second example, for instance, j learns the *higher order* information that i learned which state holds. Thus, these examples are not in conflict with the claim that in general more information may never hurt (as long as the other players do not learn anything). Moreover, they illustrate the intuitively obvious fact that it can sometimes be important te keep secret not only *what* information you have, but also the higher-order fact *that* you have secret information. The examples also show that the difference between private versus public learning is relevant not only for determining the utility of new *expected* information (questions), but even for the utility of receiving new actual information in a public or private assertion. The logical effects of learning public (versus private) information has been a major concern in recent analyses of (dynamic) epistemic logic (e.g. Fagin et. al., 1995, Gerbrandy & Groeneveld, 1997, and Baltag, Moss & Solecki, 1998) where learning higher order information is represented explicitly. Kooi (this volume) and Van Benthem (this volume) discuss how learning this kind of information also affects the updating of probabilistic information states. This section shows how these investigations are relevant for multi-agent decision situations.

4.2. Misleading information

In section 3.3 we saw that deciding by expected utility has the effect that new information can, from a first person point of view, decrease the expected utility of the (at that moment) preferred action. More serious, however, are cases where new (expected) information is considered to be useful by the agent herself, but is *disadvantageous* if we look at it from the perspective of an outside viewer who is better informed. In this respect, new information might really be *misleading*.

Consider the following decision problem discussed by Geanakoplos (1994) with an agent wondering which of $\{e_1, e_2\}$ she should perform.

[11] See, however, Bassan et al. (ms) for a result stating that, in a certain class of interactive decision situations, information will never hurt.

	e_1	e_2
t_1	-2	0
t_2	3	0
t_3	-2	0

On the assumption that the three states are equally likely, it is clear that the agent will choose action e_2 because that has, on average, a higher utility than action e_1, 0 versus $-\frac{1}{3}$. Thus, the value of choosing now is also 0, $UV(\text{Choose now}) = 0$. Suppose that t_1 is the actual situation and that our agent receives proposition $\{t_1, t_2\}$. Although this new information has a positive utility value: $UV(\{t_1, t_2\}) = \frac{1}{2}$, the new information is very *misleading* for our agent. In contrast to the initial state, after she got the information she would make the *wrong decision*: choose e_1 with a payoff of -2, instead of e_2 with a payoff of 0. Assume now that our agent asks what the actual state is to someone who cannot distinguish t_1 from t_2. Thus, this person's knowledge state can be characterized by partition $\{\{t_1, t_2\}, \{t_3\}\}$. This means that if t_1 is the actual state, he will answer with proposition $\{t_1, t_2\}$, which, again, means disadvantageous information. We can conclude that it can be a bad idea to ask someone a question who does not know the complete answer, even if he is known to be fully cooperative.[12]

However, there is another reason why we should not always act upon truthful information received from conversational participants. This second reason has everything to do with Grice's maxim of *quantity*: a cooperative speaker should not withhold relevant information! Assume, again, that our agent asks about the actual state, but that the answer that the other participant gives depends on the state: in t_1 he will assert proposition $\{t_1, t_2\}$, in t_2 he asserts $\{t_2\}$ and in t_3, $\{t_2, t_3\}$. Being an expected utility optimizer, our agent would shift her choice from e_2 to e_1 for any of these answers. But notice that although all three answers have a positive utility value, $\frac{1}{2}, 3$, and $\frac{1}{2}$ respectively, in two of the three states, she ends up making the *wrong decision*! Consider the answerer's point of view and look at the *answer strategy* S that he used: $S(t_1) = \{t_1, t_2\}, S(t_2) = \{t_2\}$, and $S(t_3) = \{t_2, t_3\}$. Assuming that $BR(p)$ denotes the best-response action of our agent after she learns proposition p, we can determine for each state t, the

[12] Or to ask a question whose complete answer does not fully resolve the decision problem. A similar point was already made by Good (1974), who concluded that a *little learning* might be dangerous.

value $U(t, BR(S(t)))$. Now we see that this value is positive only in t_2, $U(t_2, BR(p)) = U(t_2, e_1) = 3$, but negative in the other states, i.e. -2.

Above we assumed that the answerer was using strategy S. But you might wonder why he would do that. Well, suppose that independent of which state actually holds, the answerer prefers our agent to perform e_1 instead of e_2. Suppose, for instance, that the combined utility table of the answerer (first entry) and our agent can be pictured as follows:

	e_1	e_2
t_1	1,-2	0,0
t_2	1,3	0,0
t_3	1,-2	0,0

For a situation that can be modeled by the above multi-agent decision table, answer strategy S makes sense: the answerer wants our agent to shift her choice from e_2 to e_1 even in situations where this is disadvanteguous for the agent herself. However, she wouldn't always do that if he were to tell her the exact state of nature. In order to effect behavior that is favorable to himself, the answerer's strategy is to give different amounts of information in different states. If our agent takes the new information at face value, she can be *manipulated* by the answerer and will act in accordance with his, but not her own, preferences. Thus, if the preferences of the agents are not harmonically aligned, conversational participants cannot be expected to be fully cooperative, and might not say the whole truth.

But now suppose that our agent doesn't take the information she receives at face value, but instead takes into account the answer strategy employed by the other dialog participant. In that case, the utility of receiving proposition $S(t)$ in state t is not simply $U(t, BR(S(t)))$, but rather $U(t, BR(S^{-1}(S(t))))$. In that case, only proposition $S(t_2)$ has a non-zero utility, namely 3, and our agent will not act against her own preferences. And this is also expected by what we said above. In contrast to $\{S(t_1), S(t_2), S(t_3)\}$, the set $\{S^{-1}(S(t_1)), S^{-1}(S(t_2)), S^{-1}(S(t_3))\}$ does partition the state space. According to our way of determining the expected utility value of questions, this value for the latter 'question' cannot be negative, and we have just seen that this, indeed, is not the case.

5. Quality of information exchange

From the discussion in the previous section we can conclude that an agent can be manipulated by new information given by another agent in case she takes this information at face value. No manipulation is possible, however, when she knows the strategy 'behind' the assertions/answers of the other agent. Thus, the extent to which an agent can determine the answer strategy of her dialog participant is of crucial importance. Game theory is the discipline that studies, among other things, this issue. In fact, we will see in this and the following section that game theory not only discusses in which circumstances misleading information exchange is likely to occur, but also seeks to characterize situations in which *credible* information exchange cannot occur at all. A statement is credible whenever the hearer has good reasons to believe what the speaker says. The statements discussed in game theory, *threats* and *promises* in particular, typically involve actions the speaker himself intends to carry out. They are seen as strategic moves to influence the hearer's actions. The purpose of the following sections is to show that the analysis of the credibility of actions-statements can be used to study the credibility of assertions in general. At this point, these game-theoretical considerations become relevant for Gricean pragmatics, and in particular for the *maxim of Quality* which asks of the speaker not to make claims he believes are false.

5.1. Announcements in stategic games

First let's take a look at Game 1: a strategic game of pure coordination.

Game 1: coordination

	L	R
$\top$	4,4	0,0
$\bot$	0,0	2,2

A pair of simultaneous choices of actions/strategies forms a *Nash equilibrium* – the standard solution concept used in game theory to predict the actions of the agents –, if neither player can profitably deviate, given the actions of the other player. Game 1 has two pure Nash equilibria: $(\top, L)$ and $(\bot, R)$. Although the first equilibrium is preferred by both players – it is Pareto optimal – standard game theory does not single it out as the unique solution of the game. Intuitively, communication (an announcement by row-player, for instance) helps here. We can add a round of pre-play communication to the game by extending the players's strategies, and turning the strategic game into an extensive one. Row player's strategies are now not simply $\{\top, \bot\}$, but rather $\{m_\top\top, m_\top\bot, m_\bot\top, m_\bot\bot\}$, where $m_\top$ is the message that she will play $\top$. Column's strategies are $\{LL, LR, RL, RR\}$, where 'LR', for instance, means that she plays L if row-player announces $m_\top$, and R if she announces $m_\bot$. Unfortunately, even if we add to the game such a round of communication, the announcement (or *promise*) of the row-player that she will perform action $\top$ will not even eliminate the non-optimal pairs $(\top, R)$ and $(\bot, L)$ (or better $(m_\top\top, RR)$ and $(m_\top\bot, LL)$) as possible solutions of the game. They will still be counted as equilibria, because the column player can choose to ignore the announcement, and the row-player can lie about her intentions. Common wisdom, however, has it that in these circumstances, i.e., after the announcement of $m_\top$, the Pareto optimal solution will be played. Farrell (1988, 1992) and Myerson (1989) have proposed to refine the equilibrium set of such games with communication by requiring that announcements be *credible*: if the receiver believes what the signaller says, it creates for the latter an incentive to fullfil the *commitment*.[13] Row player's announcement that she will play $\top$ now has the desired effect, because her claim is credible in Farrell's and Myerson's sense. Thus, unlike our use of the Nash-equilibrium solution concept, with this refined equilibrium solution concept we can give a reason for why we communicate in the first place, even if it does not *directly* affect payoffs. Crucial for this solution is that messages have a literal, *conventional* meaning.

For an example of a game in which communication about what will be done is not in general credible, consider the *prisoners' dilemma*, where the players have the choice to cooperate or to defect:

Game 2: prisoner's dilemma:

	C	D
C	2,2	0,4
D	4,0	1,1

In this game, the assertion of the row-player saying 'I will cooperate' is not credible: regardless of whether the message is believed, row-player's best response is still to defect and column-player knows this.

[13] Farrell (1992) also gives an evolutionary motivation for why only credible commitments make sense. In fact, many of the ideas that I will discuss in the context of rational decision making are backed up with evolutionary analyses. I will leave the enormous game theoretical literature on evolutionary analyses for what it is in this paper.

Let us denote by $BR_j(e)$ the *best response* of player j to action e played by player i. Assume now that for every action e of player i there is a corresponding message m_e which says that i is going to play e. Then we can characterize as follows the *self-committing* announcements by player i about which action she is going to perform:

(1) Message m_e is **self-committing** for player i, if $e \in BR_i(BR_j(\{e\}))$

Thus, message m_e is self-committing for player i iff there is a Nash equilibrium of the game in which player i plays e. In game 2, no announcement of row-player about which action she will perform has an effect on which action column-player will choose. This is because D is the *dominant* action for row (and column) player, and only this action is self-committing.

Notice that the definition of a message as self-committing only makes reference to the relationship between sending that message and following through with that action, namely that the action must be optimal for the sender, given that the receiver receives the message in good faith and acts accordingly. The definition makes no reference to the relationship between sending a message promising one action and doing the other, e.g. that it be non-optimal. Thus, if we equate credibility of a message with its being self-committing, it is predicted that, in order to establish credibility, we don't have to bother considering alternative actions that the sender might intend to play, contrary to the message.

Aumann (1990), however, claims that this gives rise to problems in certain kinds of games and Rabin (1990) proposes a stronger condition for a message to be credible. Aumann argues that even in a game with a unique Pareto optimal Nash equilibrium – as in Rouseau's famous Stag Hunt game described below – a message saying that the sender will play her part of the Pareto optimal strategy profile cannot be taken at face value.

Game 3: Stag Hunt:

	S	R
S	5,5	0,3
R	3,0	2,2

In this game there are two (strict) equilibria: both hunting Stag, (S,S), and both hunting Rabbit, (R,R). Aumann (1990) argues that although (S,S) is Pareto optimal in this game, the message 'I will play S' is not credible. This is because both players have a strict preference over the other player's actions that does not depend on the action the player intends to play. Not only a player who intends to play S, but also a player who intends to play R has an incentive to say 'I will play S', which makes the message non-informative about the sender's intentions.

Aumann's example suggests a stronger credibility requirement, one that is indeed given by Rabin (1990). It must be the case that someone who intends to play e strictly benefits from having his message m_e believed over not having it believed. Also it must be that someone who doesn't intend to play e benefits from having m_e not believed over having it believed. Using the terminology of Farrell & Rabin (1996), a message is *self-signaling* if the speaker would want it to be believed only if it is true. Assuming that m_e is the message saying that player i will perform action e, this can be fomalized as follows:

(2) Message m_e is **self-signaling** for i if conditions 1 and 2 hold:
1 $U_i(e, e') > U_i(e, e'')$ for $e' \in BR_j(e)$ and $e'' \notin BR_j(e)$
2 $U_i(f, e'') > U_i(f, e') : \forall f \neq e \in E, e' \in BR_j(e)$ and $e'' \in BR_j(f)$

(3) A message is **credible** if it is both self-signaling and self-committing.

Thus, a message m_e is self-signaling when the sender wants it to be believed if and only if she is going to play the strategy signalled by the message. Notice that although the claim 'I will hunt Stags' is self-committing, it is not self-signaling and thus – at least according to Aumann — not credible.

The examples discussed so far are represented as strategic games where the messages correspond with the available actions. In the following sections we will think of the situations as a class of games of *incomplete information*: *signaling games.* By doing so, we take, in a sense, a broader perspective: we can now also discuss assertions not corresponding with available actions.[14] In fact, I will discuss only a particular kind of signaling game: games where messages are costless in the sense that they are payoff irrelevant.[15]

[14] In another sense, however, our perspective will be more limited, for we ignore situations where the speaker can choose a payoff relevant action himself.

[15] See van Rooy (in press) for a discussion of signaling games with payoff relevant messages. In that paper, these games are used as a foundation of Grice's maxim of *Manner*: it is explained why we use expensive signs to express marked meanings.

5.2. SIGNALING GAMES

David Lewis (1969) was the first to talk about signaling games with payoff irrelevant messages, using them to explain conventional coordination in language use. The simplest possible signaling game is formulated as follows: One individual, the signaller, is informed of the value of an uncertain parameter t (a situation, or world, but, in game theory, normally thought of as the signaller's *type*) and then chooses an action m, referred to as a message. A second individual, the receiver, observes this message (but not the value of t) and performs some action e. The payoff to each individual depends only on the value of t and the action e adopted by the receiver.

For simplicity I will assume that the strategies of sender and receiver are *functions* from types to messages, and from messages to actions, respectively. If T is the set of types, M the set of messages, and E the set of actions, a sender-strategy S is thus an element of $[T \to M]$ and a hearer-strategy R is an element of $[M \to E]$.

Now we can discuss the effect and credibility of messages in signaling games. Under the signaling game reformulation of game 1, for instance, the column player does not know the type of the row player: if her type is t she will play $\top$; and if her type is t' she will play $\bot$. To determine whether real communication is going on in such a signaling game, we have to look at the equilibria of the game.

A strategy profile $\langle S, R\rangle$ with probability function P forms a *Nash equilibrium* iff neither the sender nor the receiver can do better by unilateral deviation. That is, $\langle S, R\rangle$ and P form a Nash equilibrium iff for all $t \in T$ the following two conditions are obeyed (where $S_t = \{t' \in T : S(t') = S(t)\}$ and $U_2^*(t, S, R) = \sum_{t' \in S_t} P(t'/S_t) \times U_2(t', R(S(t'))))$:[16]

$$\begin{array}{ll} \text{(i)} & \neg\exists S' : U_1(t, R(S(t))) < U_1(t, R(S'(t))) \\ \text{(ii)} & \neg\exists R' : U_2^*(t, S, R) < U_2^*(t, S, R') \end{array}$$

Signaling games have many equilibria, but attention is mostly payed to equilibria of a particular kind. We say that communication can occur when there is a *separating equilibrium*, an equilibrium where different types of senders send different messages. These equilibria are also singled out by Lewis (1969) as *signaling systems.* Although the messages used in signaling games need not have a pre-existing, exogenously given meaning, in a separating equilibrium we can still associate a meaning

[16] Strictly speaking, this is not just a Nash equilibrium, but rather a *perfect Bayesian* equilibrium, the standard equilibrium concept for sequential, or *extensive form*, games with observable actions but *incomplete* information. In most of this paper we are only interested in separating equilibria, and we can safely ignore the probability function.

with these messages. The meaning a message can acquire must emerge solely from the strategic interaction between sender and receiver. If sender strategy S is part of a separating equilibrium, we might say that message $S(t)$ *means* t. Notice, however, that in a two-type two-message situation there might already be 2 separating equilibria. For this reason, we don't want to concentrate so much on one particular equilibrium and what the particular messages mean in this equilibrium, but rather on whether separating equilibria exist. If there is no separating equilibrium in a game with two-types, no communication is possible.

Now we want to analyze the games discussed in section 5.1 from a signaling perspective. To do so for game 1, for instance, we first turn the single game into two subgames: one (in t) where both individuals know that row-player will play $\top$, and one (in t') where both individuals know that row player will play $\bot$. Via Harsanyi's method, we then turn this set of games into a single game of incomplete information by assuming that the column player's probability function over the states is commonly known. Communication is possible in the signaling game reformulation of the game if there exists a separating equilibrium. As it turns out, this is the case for game 1, because row's preferences over column's beliefs are correlated with her intentions, i.e. with the truth. Reformulations of the prisoner's dilemma and the Stag Hunt game show that, when this condition is given up, costless communication can not occur. In the Stag Hunt game, for instance, the speaker wants the hearer to think that she is of type t that hunts Stag: she wants the hearer to hunt Stag no matter what her type is.

Stag Hunt:

	S	R
t	5,5	0,3
t'	3,0	2,2

But this means that there can be no communication, i.e., there is no separating equilibrium because senders of type t' don't want to distinguish themselves from senders of type t.

For the simple cases that we have considered so far – strategic situations with two types of senders and where the receiver can choose between two actions – we can show that credible communication is possible in very specific situations only (see Gibbons, 1992). Consider the following abstract table:

two-type, two-action:

	e_H	e_L
t_H	$x, 1$	$z, 0$
t_L	$y, 0$	$w, 1$

It is easy to see that in this two-type, two-action situation, communication (i.e. a separating equilibrium) is possible only in case $x \geq z$ and $y \leq w$. We can check this by looking at the other possible cases: (i) if $z > x$ and $y > w$ the preferences are strictly opposed. No communication is possible now, because t_H would like the hearer to believe that his type is t_L, and the other way around for a sender of type t_L. Thus the signaling game will have no Nash equilibrium; (ii) if $x > z$ and $y > w$ (or if $z > x$ and $w > y$) *both* types of sender prefer the same action of the receiver: in our example both prefer action e_H to action e_L. No communication will take place in this case either, because both players want the receiver to believe that her type is t_H. This game is the same as the Stag Hunt game, in which we already saw that no communication was possible. We can conclude that in the simplest two-type, two-action situations, costless communication is possible only in case the preferences are perfectly aligned.[17,18]

5.3. Credibility: conventional and speaker's meaning

Until now we have assumed that a message used by a sender in a signaling game is just an element of M, but we have not determined what the elements of M are or mean. In fact, we didn't have to: our analysis was not dependent on the 'meanings' of the messages, but only on whether two messages are the same or not. But, of course, we could assume with Farrell, Myerson, and Rabin that messages have a (pre-existing) conventional meaning. Suppose that '$[[\cdot]]$' is an exogenously given interpretation function mapping messages to their meanings, i.e. propositions, subsets of T. In that case we might make for each $t \in T$

[17] Watson (1996) notes, however, that this conclusion is based on the assumption that only the receiver is uninformed. Something different happens if the receiver also has a type, which then the sender is uninformed of. In that case the receiver can pretend that the preferences are more aligned, and the sender will give more information than he would if he had perfect information. Thus, in a fully revealing equilibrium, the receiver may prefer the sender to be confused.

[18] This result, and others discussed in this paper, suggests that in zero-sum games, no credible communication is possible between rational agents. This seems to be bad news for Merin (1999), who bases his analysis on the hypothesis that language users are crucially engaged in zero-sum games. Fortunately, almost none of the results he employs to account for specific phenomena crucially makes use of this hypothesis.

the following distinction between conventional meaning and speaker's meaning (see Grice, 1967) if the speaker adopts sender strategy S.

(4) a. The *conventional meaning* of $S(t)$ is $[[S(t)]]$

b. The *speaker's meaning* of $S(t)$ is $S_t = \{t' \in T : S(t') = S(t)\}$

The intuition behind the notion of speaker's meaning is that if a player sends different messages at different states, a hearer should be able to differentiate these states. In a state t, a receiver not only learns the information content of communicated message $[[S(t)]]$, but also takes into account what the sender would have revealed in other states.[19] Thus, when the receiver knows the communication strategy of the sender, she learns not only the literal content of the message, but also something more. Notice that the intersection $[[S(t)]] \cap S_t$ can, in principle, be empty. Thus, if we call $[[S(t)]] \cap S_t$ the *communicated meaning* of $S(t)$, we are basing this notion of meaning on a convention of *truthfulness* (cf. Lewis, 1969). Assuming that senders know their type,[20] this means that truthfulness depends on the notion of *credibility* discussed earlier. In particular, the expectation that $[[S(t)]] \cap S_t \neq \emptyset$ depends on the assumption that all messages that are sent are self-signaling.

In terms of exogenously given conventional meanings of messages, we can formalize the self-signaling condition for signaling games that we discussed earlier for strategic form games.[21] This formalization is based on the (temporary) simplifying assumption that messages are used to uniquely single out the type of the sender. We say, for instance, that m_t is the message stating that the sender is of type t.

(5) Message m_t is **self-signaling** for i if conditions 1 and 2 hold:
1 $U_i(t,e) > U_i(t,e')$ for $e \in BR_j(t)$ and $e' \notin BR_j(t)$
2 $U_i(t',e') > U_i(t',e) : \forall t' \neq t \in T, e \in BR_j(t)$ and $e' \in BR_j(t')$

[19] Note that, in a fully separating equilibrium, S_t will always be $\{t\}$. Later, however, we will discuss situations where S_t is not necessarily so trivial.

[20] Until now we have assumed that the sender knows which state she is in. However, we might relax this assumption. Suppose that K is an *epistemic* accessibility relation between states and that the sender knows in t that she is in one of the states $K(t)$. Then the speaker's strategy depends not on the state she is in, but rather on what she knows: for all $t, t' \in T$, if $K(t) = K(t')$, then $S(K(t)) = S(K(t'))$. In computer science it is commonly assumed that a *protocol* is a function from information states to actions (cf. Parikh & Ramanujam, this volume). Thus, we might think of communication protocols as special kinds of sender strategies.

[21] Notice that the self-committing condition doesn't make much sense now, because the row-player has no payoff relevant action to choose anymore.

The first part of definition (5) says that if player i is of type t, then she prefers m_t to be believed so that player j plays her optimal action in t and not some other action e'. The second condition states that player i wants her message to be believed (the other player plays a best response against it) only if she is of the type announced in the message. Let us say that in signaling games a message is credible if and only if it is self-signaling. We see that, in a two-type situation, we have a separating equilibrium iff there is at least one credible message m_t.

With Lewis (1969) we can say that, in a particular game-theoretical situation, the speaker *intends* to communicate S_t with his use of signal $S(t)$. Note that this can be accounted for without assuming any mention of the (conventional) meaning of the signal at all. It seems not unnatural to call what the speaker intends to communicate with the use of a signal its 'speaker's meaning', discussed a lot in the philosophical literature. But how can the hearer recognize what the speaker meant? If she assumes that the sender is using strategy S because it is part of the (unique) most salient Nash equilibrium then she can calculate S_t. And indeed, for a signal to have a meaning, the hearer has to (be able to) *recognize* what the speaker intends.[22]

Earlier in this paper I discussed the semantics and pragmatics of natural language with respect to primacy. We have seen that the semantic notion of entailment can be reduced, to a large extent, to the pragmatic notion of usefulness. In this section, however, we suddenly assumed the existence of an exogenously given semantic interpretation function, suggesting that the semantic notion of *conventional meaning* cannot be reduced to the pragmatic notion of *speaker's meaning*. However, Lewis's (1969) work on signaling games strongly suggests the logical priority of the latter notion: speaker and hearer can coordinate their behavior with the help of signals without them having pre-existing conventional meanings. In fact, Lewis thinks of the conventional meaning of expressions as the result of solutions of recurrent coordination problems. Typical (recurrent) coordination problems out of which conventional meanings arise are as simple as 'for object x in the world, how can a speaker talk about it such that the hearer can understand?'. This suggests that the conventional meaning of an expression must be relatively *general*, and leaves the *actual* interpretation of the use of the expression in a concrete situation *underspecified*. In fact, as stressed by Clark (1996) and others, this is yet another argument in favor of the logical priority of speaker's meaning: the actual interpretation depends not only on the

[22] See Prashant Parikh (2001) for a more extensive discussion of the notion of 'speaker's meaning' from a game-theoretical perspective.

conventional meaning of an expression, but also on intentions, beliefs, and preferences of the participants involved in the conversation.

6. Quantity of information exchange

In the previous section we only discussed situations with two types. Communication is possible in such signaling games just in case there is a separating equilibrium. In situations with more types and more actions, however, we have more interesting alternatives. In these cases, communication is possible even when not every type sends a different message. For these situations the following question arises: With what degree of *fine-grainedness* can a speaker credibly reveal her type?

In an important article, Crawford & Sobel (1982) show that the amount of credible communication depends on how far the preferences of the participants are aligned. They show that when the preferences are more aligned, more communication can occur through costless signaling. They construct utility functions for sender and receiver such that the equilibria in such games are *partition equilibria*; i.e., the type set can be partitioned into a finite number of sets such that types belonging to the same set send a common message and receive the same action. The more fine-grained this partition is, the more communication is possible, and the fine-grainedness of the partition depends on the extent to which the preferences are aligned.

Crawford & Sobel work with continuous type, message, and action spaces. The idea behind their analysis does not depend on this, however. Consider first the following two-type, two action game.

Game 4:

	e_1	e_2
t_1	1,1	0,0
t_2	0,0	1,1

This game obviously gives rise to the partition equilibrium $\{\{t_1\}, \{t_2\}\}$ with best replies $\{\{e_1\}, \{e_2\}\}$.

Now consider a game with more than 2 actions involved:

Game 5:

	e_1	e_2	e_3
t_1	3,3	1,0	2,2
t_2	1,0	0,3	2,2

Suppose that column-player takes both types to be equally likely. In that case he will choose e_3, because that has, on average, the highest utility. Notice that if the speaker is of type t_1 and can convince the column-player that she is of type t_1, action e_1 will be chosen which is favourable for both. Unfortunately, however, the claim 'I am of type t_1' cannot be believed: although the receiver's best response to this claim is action e_1, BR('I am t_1') $= e_1$, which is the one that the sender prefers, someone of type t_2 doesn't want to distinguish herself from someone of type t_1 (she wants column-player to be unclear whether row-player is of type t_1 or of type t_2, in which case she gets a payoff of 2, which is the best she can expect) and thus would send the same message. If the sender of type t_2 reveals her type, the receiver would play e_2, which is the worst for the sender. Thus, there can be no information transmission going on: nobody would say 'I am t_2', and the receiver cannot trust the claim 'I am t_1'. As a result, they will send the same message in equilbrium. An equilibrium in which all individuals send the same message is called a *pooling equilibrium.* Notice that this case is somewhat different from the ones talked about before, because now a third action e_3 is in play. This action will be chosen by the receiver.

The following game is somewhat more complex.

Game 6:

	e_1	e_2	e_3	e_4
t_1	4,4	1,1	0,0	3,3
t_2	1,1	1,4	0,0	3,3
t_3	0,0	0,0	3,4	0,0

Just as in the previous game, here type t_1 individuals cannot distinguish themselves from type t_2 individuals through costless signaling. Recall our definition (5) of self-signaling given earlier:

(5) Message m_t is **self-signaling** for i if conditions 1 and 2 hold:
1 $U_i(t,e) > U_i(t,e')$ for $e \in BR_j(t)$ and $e' \notin BR_j(t)$
2 $U_i(t',e') > U_i(t',e):\ \forall t' \neq t \in T, e \in BR_j(t)$ and $e' \in BR_j(t')$

Notice that m_{t1} is not self-signaling for the second reason, because there are $t \neq t_1$ and $e \in BR_j(t)$, namely t_2 and e_2, such that $U_i(t_2,e_2) \leq U_i(t_2,e_1)$. Message m_{t2}, on the other hand, is not self-signaling for the first reason: $U_i(t_2, BR_j(t_2)) = U_i(t_2,e_2) < U_i(t_2,e_4)$. In fact, only m_{t3} is self-signaling according to (5). However, individuals of type t_1 and t_2 should intuitively be able to distinguish themselves form t_3 -type individuals. To account for this intuition, we have to generalize the above definition. Instead of talking about types, we should rather talk about *sets* of types (in the following, $S(t)$ stands for $[[S(t)]]$):[23]

(6) Message $S(t)$ is **self-signaling** for i if conditions 1 and 2 hold:
1 $\forall t' \in S_t : U_i(t', e) > U_i(t', e')$ if $e \in BR_j(S(t))$ & $e' \notin BR_j(S(t))$
2 $\forall t' \notin S_t : U_i(t', e') > U_i(t', e)$ if $e \in BR_j(S(t))$ & $e' \in BR_j(S(t'))$

Now we see that the message *I am either of type* t_1 *or* t_2 is self-signaling: if $i \in \{1, 2, 4\}$, then (i) $U_1(t_1, e_i) > U_1(t_1, e_3)$ and $U_1(t_2, e_i) > U_1(t_2, e_3)$, and (ii) $U_1(t_3, e_3) > U(t_3, e_i)$.

Notice that the above definition depends on strategy S. Let us now say that a **game** is *self-signaling* with respect to sender-strategy S if each $t \in T : \;\; S(t)$ is self-signaling. On our assumption that S is a *function* from types to messages, this definition of self-signaling gives rise to *partitions* of both the types and the actions in $BR(T)$: (i) although $[[S(t)]]$ might have a non-empty intersection with $[[S(t')]]$ if $t \neq t'$, it will be the case that S_t and $S_{t'}$ are either the same or have no element in common. Thus, S gives rise to partition $\{S_t : t \in T\}$; (ii) the conditions on self-signaling have the effect that $\forall t, t' : S(t) \neq S(t') \rightarrow BR([[S(t)]]) \cap BR([[S(t')]]) = \emptyset$.

In the last example we have a (non-trivial) partition equilibrium – $\{S_t : t \in T\} = \{\{t_1, t_2\}, \{t_3\}\}$ with as the set of corresponding best replies $\{\{e_1, e_2, e_4\}, \{e_3\}\}$ –, but one that is *not* completely separating. In equilibrium, it makes no sense for individuals of type t_1 and type t_2 to send different messages, although they send a different message than an individual of type t_3. An equilibrium of this type is called a *partial pooling equilibrium.*

Notice that there might be more than one game and sender-strategy pair that is self-signaling. In game 6, for instance, also sender-strategy S' that assigns to each type the same message satisfies (trivially) the above condition. (This is in correspondence with the fact that costless signaling games always have a pooling equilibrium.) It seems obvious, however, that sender-strategy S is in this game preferred to strategy S': more information is transmitted via this strategy. This gives rise to the following idea: the expected sender-strategy is the self-signaling one that gives rise to the *finest partition* of the types.[24]

[23] Where $BR_j([[S(t))]])$ is the set of actions that are best responses given *some* beliefs consistent with $[[S(t)]]$. Thus, we look not at one, but rather at a set of Nash equilibria now. Each of e_1, e_2 and e_4 are in $BR_j(\{t_1, t_2\})$, for instance, because there is a probability function that makes e_1 the optimal response given $\{t_1, t_2\}$ (if t_1 is taken to be more likely), but also ones that makes e_2 or e_4 optimal.

[24] I believe that this comes down to, and is certainly inspired by, the PCI condition of Blume et al. (2001). However, they don't relate it to self-signaling. The reason

The question of how much communication can be expected in signaling games can now be answered by determining how fine-grained the partition-equilibrium will be. Just as in Crawford & Sobel (1982), the fine-grainedness of the partition depends on how far the preferences of sender and receiver are aligned. Note that this affects the decision which question to ask and the expectation which answer to anticipate: it doesn't make much sense to ask a question that gives rise to a finer partition than the one that self-signaling induces.

7. Conclusion

In the first part of this paper we discussed the value, or relevance, of questions and assertions by relating them to decision problems. In the second part we investigated under which circumstances, and with how much credibility, information can flow between the participants of a dialog. Throughout the paper I have (more or less explicitly) related these decision and game theoretical issues with Gricean concerns for natural language analysis. First, we have shown how the knowledge of strategies of conversational agents in signaling games can be used to determine what is meant by the speaker of an expression with a conventionally non-existing, or underspecified, meaning. Second, by making use of standard decision theoretic methods, it became clear how we can determine the quantity of relevant information given in an assertion and asked for in a question. This enables us to give a much more accurate formalization of Grice's maxims of quantity and relevance than normally assumed. But perhaps the most important result of this paper was reached in the final sections: it showed the limitation of Grice's most basic assumption: his cooperative principle. Where Grice (1967) appeals to rationality to motivate his assumption that speakers are fully cooperative language users, our discussion rather suggests that, as far as rationality is concerned, they can be expected to be cooperative only in so far as their preferences are (known to be) aligned. Notice that especially with respect to the second and third issue, decision and game theory are found to be useful not only to think of language use from a broader and conceptually more appealing perspective than usually, but also because some decision and game theoretical results (e.g. those of Blackwell and Crawford & Sobel) are shown to be of direct relevance to semantic/pragmatic concerns.

This paper does not stand alone. It is part of the growing decision and game-theoretical literature on language use (eg. P. Parikh (1991,

behind this, I guess, is that they want to state their PCI condition independently of the assumption that messages have a pre-existing meaning.

2001), R. Parikh (1994), Merin (1999), Asher et. al. (2001), and van Rooy (2003, in press)) which promises to shed new light on existing problems. In contrast to this more general paper, in most of the others, decision and game theory are used to analyze some more particular phenomena. These papers make use of a shared set of analytical tools. More importantly, however, they also share an important methodological assumption: the primacy of pragmatics with respect to natural language semantics.

Acknowledgements

This research for this paper has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences (KNAW). I would like to thank Boudewijn de Bruin, Balder ten Cate, Peter Grünwald, Katrin Schulz, Frank Veltman, but especially Johan van Benthem for comments and discussion. Special thanks to William Rose for correcting my English and further discussion.

References

Asher, N. I. Sher & M. Williams (2001), Game theoretical foundations for Gricean constraints, *Proceedings 13th Amsterdam Colloquium*, 31-36.

Aumann, R. (1990), Nash equilibria are not self-enforcing, In: J. Gabszewicz et al (eds.), *Economic Decision-Making: Games. Econometrics and Optimization*, Elsevier, Amsterdam, 201-206.

Baltag, A., L. Moss & S. Solecki (1998), The logic of public announcements and common knowledge, In *TARK 98*.

Bassan, B., O. Gossner, M. Scarsini & S. Zamir (ms), Positive value of information in games, manuscript.

Benthem, J. van (this volume), Conditional probability and update logic, *JoLLI*.

Blume, A., D. DeJong, Y. Kim & G. Sprinkle (2001), Evolution of communication with partial common interest, *Games and Economical Behavior*, **37**: 79-120.

Clark, H. (1996), *Using Language*, Cambridge University Press, Cambridge.

Crawford, V. & J. Sobel (1982), Strategic information transmission, *Econometrica*, **50**: 1431-51.

Fagin, R., J. Halpern, Y. Moses & M. Vardi (1995), *Reasoning about Knowledge*, MIT Press, Cambridge, Massachusetts.

Farrell, J. (1988), Communication, coordination and Nash equilibrium, *Economic Letters*, **27**: 209-214.

Farrell, J. (1993), Meaning and credibility in cheap-talk games, *Games and Economic Behavior*, **5**: 514-531.

Farrell, J. & M. Rabin (1996), Cheap talk, *Journal of Econ. Perspectives*, **10**: 103-18.

Geanakoplos, J. (1994), Common knowledge, In: R. Aumann & S. Hart (eds.), *Handbook of Game Theory*, Elsevier, Leiden, 1438-1496.

Gerbrandy, J. & W. Groeneveld (1997), Reasoning about information change, *Journal of Logic, Language, and Information*, **6**: 147-196.

Gibbons, R. (1992), *A primer in Game Theory*, Harvester Wheatsheaf, New York.
Good, I.J. (1974), A little learning can be dangerous, *British Journal for the Philosophy of Science*, **25**: 340-342.
Grice, H.P. (1957), Meaning, *Philosophical Review*, **66**, 377-388.
Grice, H. P. (1967), Logic and Conversation, typescript. In: P. Grice (1989), *Studies in the Way of Worlds*, Harvard University Press, Cambridge, 22-40.
Groenendijk, J. & M. Stokhof (1984), *Studies in the Semantics of Questions and the Pragmatics of Answers*, Ph.D. thesis, University of Amsterdam.
Hirshleifer J. & J.G. Riley, (1992), *The Analytics of Uncertainty and Information*, Cambridge University Press, Cambridge.
Kooi, B. (this volume), Probabilistic dynamic epistemic logic, *JoLLI.*
Lewis, D. (1969), *Convention*, Harvard University Press, Cambridge Massachusetts.
Luce, R.D. & H. Raiffa (1957), *Games and Decision*, John Wiley, New York.
Merin, A. (1999), Information, relevance, and social decisionmaking, In: L. Moss et al. (eds.), *Logic, Language, and Computation. Vol. 2*, CSLI, Stanford.
Myerson, R. (1989), Credible negotiation statements and coherent plans, *Journal of Economic Theory*, **48**: 264-303.
Neyman (1991), The positive value of information, *Games and Economic Behavior*, **3**: 350-355.
Osborne, M.J. & A. Rubinstein (1994), *A Course in Game Theory*, MIT Press, Cambridge, Massachusetts.
Parikh, P. (1991), Communication and strategic inference, *Linguistics and Philosophy*, **14**: 473-513.
Parikh, P. (2001), *The use of Language*, CSLI Publications, Stanford.
Parikh, R. (1994), Vagueness and utility: The semantics of common nouns, *Linguistics and Philosophy*, **17**: 521-535.
Parikh, R. & R. Ramanujam (this volume), A knowledge based semantics of messages, *Journal of Logic, Language and Information.*
Rabin, M. (1990), Communication between rational agents, *Journal of Economic-Theory*, **51**: 144-170.
Raiffa, H. & R. Schlaifer (1961), *Applied Statistical Decision Theory*, MIT Press, Cambridge, Massachusetts.
Rooy, R. van (2003), Questioning to resolve decision problems, *Linguistics and Philosophy*, to appear.
Rooy, R. van (in press), Signaling games select Horn strategies, *Ling. and Philosophy.*
Rooy, R. van (to appear), Utility, informativity and protocols, *Journal of Philosophical Logic.*
Savage, L.J. (1954), *The Foundations of Statistics*, Wiley, New York.
Shannon, C. (1948), The Mathematical Theory of Communication, *Bell System Technical Journal*, **27**: 379-423 and 623-656.
Szaniawski, K. (1967), The value of perfect information, *Synthese*, **17**: 408-424.
Wald, A. (1950), *Statistical Decision Functions*, John Wiley, New York.
Watson, J. (1996), Information transmission when the informed party is confused, *Games and Economical Behavior*, **12**: 143-161.

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Nieuwe Doelenstraat 15
1012 CP Amsterdam
vanrooy@hum.uva.nl