Chapter 1

CONVERSATIONAL IMPLICATURES AND COMMUNICATION THEORY

Robert van Rooy*
Institute for Logic, Language and Computation
University of Amsterdam
Nieuwe Doelenstraat 15, 1012 CP Amsterdam
vanrooy@hum.uva.nl

Abstract According to standard pragmatics, we should account for conversational implicatures in terms of Grice's (1975) maxims of conversation. Neo-Griceans like Atlas & Levinson (1981) and Horn (1984) seek to reduce those maxims to the so-called Q and I-principles. In this paper I want to argue that (i) there are major problems for reducing Gricean pragmatics to these two principles, and (ii) that, in fact, we'd better account for implicatures in terms of the principles of (a) *optimal relevance* and (b) *optimal coding.* To formulate both, I will make use of Shannon's (1948) mathematical theory of *communication.*

1. Introduction

Natural language is flexible in the sense that a single message can convey different semantic contents in different contexts. And indeed, recent trends in semantics (e.g. optimality theoretic semantics) suggest that the actual interpretation of an utterance is highly *underspecified* by the conventional meanings of the sentence that is used. This requires, however, that language users have robust ways to resolve the under-

*This paper reports about work in progress on the use of general theories of action for the semantic/pragmatic analysis of natural language. This paper is an extension of my original contribution presented at the 2nd SIGdial workshop. I have given a similar presentation at a Stanford workshop on Logic and Information in 2000. I would like to thank the participants of those workshops and the anonymous reviewers for their comments and suggestions. This research has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences (KNAW), which is gratefully acknowledged.

specification and/or ambiguity. In this paper I will discuss two ways of doing this. First, one where the *particular conversational* situation is important; second, one which depends on more *general conventions*.

2. Particularized Conversational Implicatures

2.1 The *Q* and *I* principle

Neo-Gricean pragmatics (Atlas & Levinson, 1981; Horn, 1984; Levinson 2000) seeks to reduce Grice's maxims of conversation to the so-called Q and I principles. Both are used to account for many conversational implicatures. The Q-principle (implementing Grice's first maxim of Quantity) advises the speaker to say as much as he can to fulfill his communicative goals, while the I-principle (implementing Grice's other maxims, except for quality) advises the speaker to say no more than he must to fulfill these goals. Both principles help to strengthen what is communicated by a sentence. The Q-principle induces inferences from the use of one expression to the assumption that the speaker did not intend to communicate a contrasting, and informationally stronger, one. This principle is thus essentially metalinguistic in kind, and accounts for both *scalar* and *clausal* implicatures. It allows us, for instance, to conclude from 'John ate *some* of the cookies' to 'John didn't eat *all* of the cookies' (scalar implicature), and from 'A or B' to 'A or B, but not both' (clausal + scalar implicature). The I-principle allows us to infer from the use of an expression to its most informative or stereotypical interpretation. It is used, for instance, to enrich the interpretation of a conjunction to a temporal sequential, or causal, relation, and it allows us to interpret a conditional like 'John walks, if Mary walks' as the biconditional 'John walks if and only Mary walks'.

2.2 Problems for the *Q* and *I* principles

Although the Q and I principles are intuitively appealing, they give rise to a number of conceptual and empirical problems. They both *under*- and *over*generate.

2.2.1 Too general. Let's start with some cases where it is predicted that Q-implicatures arise, although in fact they don't. First, at least when implemented as Gazdar (1979) did, we can derive from the existential 'Someone is sick' as a Q-implicature that (the speaker knows that) a is not sick, for any individual a. Second, on the assumption that scales are defined in terms of entailment, it is predicted that we can infer from 'B, if A' to the conclusion that it is not the case that the stronger 'B if and only if A' holds, although in a lot of situations this is exactly what we can conclude. Third, on the same assumption, it is incorrectly predicted that we can infer 'not *regret* A' from 'know A'. Gazdar, Horn, Levinson and others have argued that these problems can be prevented by cancellation by clausal implicature and/or weakening the force of Q-implicatures from know-not to not-know (for the first problem), and by putting constraints on what counts as contrastive expressions: contrastive expressions must be lexical items (second problem) and must have the same presuppositions (for the third). Although it can be argued that for the biconditional interpretation this – somewhat ad hoc – solution solves the second problem, Gazdar (1979) argued that the constraints doesn't solve the third one. Moreover, the most serious problematic cases where Q-implicatures overgenerate cannot be explained away in this way: The Horn/Gazdar/Levinson/Atlas analysis of Q-implicatures as *generalized* conversational implicatures (PCIs) triggered solely by lexical expressions cannot explain why from A's answer 'John has 2 children' to Q's question 'Who has 2 children?' the implicature 'John has only 2 children' does not even arise as a default (cf. van Kuppevelt). This latter example seems to suggest that these so-called Q-implicatures are, after all, dependent on the conversational situation, in particular on the question being asked (e.g. Hirschberg, van Kuppevelt). Proponents of the Q and I pragmatics (Horn, Levinson), followed by Matsumoto (1995), argue that in such *particular* conversational situations the *generalized* conversational implicature is *cancelled*, for reasons of relevance: The answer is already *informative* enough for the purpose of the conversation. I will argue, however, that informativity is, in general, not the crucial issue, and that it is much more natural to assume that – for reasons of relevance in this particular situation – the (potential) implicature does not even arise.

2.2.2 Not general enough. Not only does the standard analysis of Q-implicatures overgeneralize, it also doesn't seem to be general enough. First, as discussed extensively by Hirschberg (1985), the standard analysis is of no help to account for certain examples that intuitively should be analyzed as scalar implicatures. If Mary's potential new boss asks her at her job-interview whether she speaks French, and she answers by saying 'My sister does', he can conclude that Mary herself does not. The standard analysis fails to account for this, because (a) scalar implicatures are all analyzed in terms of the Q-principle, (b) the Q-principle is stated in terms of *informativity*, but (c) the proposition that Mary speaks French is not more *informative* (i.e. entails) than the proposition that her sister does. This example suggests (i) that scalar implicatures should not exclusively be accounted for in terms of informativity, and (ii) that just like in the previous example, also here the relevant implicature crucially depends on the conversational situation (i.e. the beliefs and preferences of the agents involved). Second, as discussed by McCawley (1993), the implicatures generated by the ⟨and, or⟩ scale cannot account for the fact that a sentence of the form '*A or B or C*' gives rise to the inference that only one of the three is true. A final example where the standard analysis of Q-implicatures isn't general enough was discussed by Groenendijk & Stokhof (1984). They observe that when A answers Q's question 'Who comes?' by saying 'Peter comes', we typically interpret the answer as being exhaustive. That is, we interpret A's answer as 'Only Peter comes'. They claim that this kind of inference should intuitively be accounted for in terms of Grice's maxim of Quantity (as a Q-implicature), but note that the standard implementation does not predict the exhaustivity of the answer. Still, it seems that the exhaustive interpretation of the answer should be derived by Gricean pragmatics on the assumption that answers are as informative as the question requires.

I conclude that the scales relevant for the implicatures depend on the conversational situation (i.e. question asked) and the beliefs and preferences of the agents involved is in correspondence with Hirschberg's claim that scales are dependent on context. However, we would like to say something more; we would also like to say *how* the relevant scale depends on the question asked and the relevant beliefs and desires.

2.3 Relevance

In this respect, important progress has been made recently by Merin (1997). Following the lead of Anscombre & Ducrot (1983), Merin argues that scales should be defined not in terms of informativity, but rather in terms of a notion of *relevance*. The relevance of a proposition is determined in terms of the *argumentative force* the proposition would have in that particular conversational situation. The relevance of an assertion is then defined in information/decision/game theoretical terms, based on the assumption that the participants of the conversation have strictly opposing preferences, i.e. that the participants play a zero-sum game.

Although Merin convincingly shows that some scalar implicatures (in particular the Hirschberg examples) can be accounted for appropriately on the assumption that players argue for particular hypotheses, and that their contribution should be interpreted in the most relevant way (i.e. strongest argument), it is intuitively clear that not all conversations can, and should, be modeled as zero-sum games. It makes little sense, for instance, to assume that the exhaustive interpretation of 'John has 2 children' as answer to the question 'How many children does John have?' can be explained in terms of opposing preferences between questioner and answerer, for the latter typically cooperates with the former. What is called for, then, is a *generalization* of Merin's notion of relevance that also measures the relevance of propositions in *cooperative* conversational situations. It seems only natural, on the assumption that speakers are relevance optimizers, that once we can define such a measure, not only the typical Q-implicatures can be accounted for in terms of relevance, but also the I-implicatures from conditional to biconditional, and Groenendijk & Stokhof's (1984) observation that answers are normally interpreted in an exhaustive way. As we will see in the next section, Groenendijk & Stokhof (1984) show that almost all typical Q-implicatures can be analyzed alternatively in terms of their explicit exhaustivity-operator, without giving rise to the above discussed overgeneralizations, when the clause that gives rise to the implicature is used as an answer to a question.

2.4 Exhaustified answers

Groenendijk & Stokhof (1984) propose to account for the intuition that answer *Peter comes* to question *Who comes?* should normally be read exhaustively by introducing an explicit exhaustivity operator that is applied to answers and the abstracts underlying the questions to derive the exhaustive interpretation.

$$\underline{exh} \quad = \quad \lambda R \lambda P[R(P) \wedge \neg\exists P'[R(P') \wedge P \neq P' \wedge \forall x[P'(x) \rightarrow P(x)]]]$$

This exhaustivity operator accounts for many of the implicatures traditionally accounted for in terms of Grice's maxim of quantity. First, it obviously accounts for the fact that when *Who comes?* is answered by *John* we conclude that *only* John comes. Second, when answer *A man* is given we can conclude that not all men come, an implicature standardly triggered by the ⟨all, some⟩ scale. Note that this analysis, in distinction with the standard analysis of scalar implicatures, works also well when more than one item gives rise to an implicature. From the exhaustive interpretation of the term *Some of the bacon and some of the eggs* given as answer to the question *What did Mary ate?* we can conclude that Mary didn't eat all of the bacon, and that she didn't eat all of the eggs, just like we should.

Notice that our exhaustification analysis not only predicts intuitions standardly accounted for in terms of the Q principle; also some I-implicatures are accounted for. If the question is *Who quacks?* the answer *Every duck quacks* is predicted to mean that every quacker is a duck. Horn (2000) calls this inference *conversion* and explicitly proposes to account for it in terms of the *I*-principle.

Similarly, if we allow for explicit quantification over worlds, we can account for the inference from (1b) to (1c), when the former is given as answer to (1a):

(1) a. Q: Did John walk?

b. A: *If* Mary talked.

c. John walked *iff* Mary talked.

We assume with Groenendijk & Stokhof (1984) that the property underlying (1a) is $\lambda w.Walk(j)(w)$, and that answer (1b) should be represented by $\lambda p.\lambda w[Talk(m)(w) \rightarrow p(w)]$ which after exhaustification means that Mary talked iff John walked.

Our approach also predicts that (2a) should be read as (2b) when the color of the flag is at issue.

(2) a. The flag is red.

b. The flag is all red.

This inference is normally (e.g. Atlas & Levinson, 1981) accounted for by assuming that (2a) should be interpreted as informative as possible. But then it should be explained why in certain circumstances the inference is absent. Suppose 3 flags are mutually known by us to be all white except for a small patch of color (being red, yellow, or green). Because the color of the patch distinguishes the flags, if I ask you to identify the flag you hold behind your back, your answer (2a) satisfies me, and I do not imply that (2b) is true. The standard analysis has to assume that in these cases the triggered generalized implicature are *cancelled*. An exhaustivity-based analysis of scalar implicatures as suggested by Groenendijk & Stokhof (1994) and more explicitly proposed by van Kuppevelt (1996) doesn't have to assume that scalar implicatures can be cancelled for reasons of relevance, because now the notion of scalar implicature is made *topic*-dependent. For Groenendijk & Stokhof, topics are explicitly asked questions. Van Kuppevelt (1996) proposes that exhaustification can also take place with expressions in *comment* position which can be part of assertions given as answers to *implicit questions*.[1] Now, we don't even generate the implicature because we can assume that the implicit question was something like *What is the color of the patch?*

Indeed, a topic-dependent analysis of 'scalar' implicatures prevents us from triggering implicatures to be *cancelled* later for reasons of relevance.[2] Consider van Kuppevelt's example again:

(3) a. Q: Who has 2 children?

b. A: John has 2 children.

c. John doesn't have more than 2 children.

Instead of saying that (3b) triggers the potential implicature (3c) that is cancelled when the former is given as answer to question (3a), our analysis predicts that the implicature is not even triggered, *because* (3b) completely answers (3a).

A similar analysis can be given for the fact that a disjunctive sentence sometimes gets an exclusive reading and sometimes not. Consider sentence (4).

(4) The cookies or the chocolates are in the box.

If we assume that this sentence is given as answer to the question 'Are the cookies in the box?', the application of exhaustivity results in the exclusive reading. However, this analysis does not have the result that a disjunctive sentence should always have the exclusive reading. In particular this is rightly predicted not to be the case if (4) is given as an (exhaustive) polar answer to (5).

(5) Q: Are the cookies or the chocolates in the box?

A: Yes, the cookies *or* the chocolates are in the box.

Something similar is the case with conditional answers. Also after exhaustification they don't get a bi-conditional interpretation when they are used as complete answers to polar questions:

(6) Q: Did John walk, if Mary talked?

A: Yes, John walked *if* Mary talked.

2.5 Relevance and Exhaustivity

In the previous section we have seen that many so-called 'quantity' implicatures triggered by sentences can be accounted for by assuming that these sentences should be interpreted as exhaustive answers to questions. However, we would like to say something more; we would also like to give an independent motivation for *why* answers should normally be interpreted exhaustively. Notice that Groenendijk & Stokhof's (1984) stipulation that answers should always be interpreted exhaustively would not only be *ad hoc*, it would also give rise to *counterexamples*. Most importantly, it would predict incorrectly for so-called mention-some questions. Sometimes an assertion intuitively answers a question completely without being read exhaustively. To illustrate, when I ask you (7a) and you answer by saying (7b), I am satisfied, although I don't interpret your answer as claiming that this is the *only* place where I can buy an Italian newspaper.

(7) a. Where can I buy an Italian newspaper?

b. Around the corner.

2.6 Topic dependent relevance

In cooperative dialogues the relevance of communicative acts can be determined with respect to *decision problems* (cf. van Rooy (2001). Using Shannon's (1948) communication theory we can model these decision problems by partitions of the logical space –, i.e., the semantic *questions* of Groenendijk & Stokhof (1984). One proposition will then be more relevant than another when it helps more to resolve the question. Intuitively, we would like to say that assertions are relevant with respect to this decision problem if the decision is easier to make after an assertion is learned. But to account for this, we have to *measure* the difficulty of the decision. A standard way to do this is in terms of *entropy.*

Given a probability function P, we can define the *entropy* of decision problem Q as follows:

$$E(Q) \quad = \quad \sum_{q \in Q} P(q) \times -log_2 P(q)$$

If Q has 8 answers, $E(Q)$ is *maximal* exactly when all answers have equal probability to be true: $\frac{1}{8}$. The surprise value of each of its answers A is then $log_2 \frac{1}{P(A)} = \inf(A) = 3$, and so its *average* surprise value, i.e. its entropy, is 3 too. For yes/no questions, for instance, this means that the entropy is maximal in case both answers have a probability of 0.5. When our agent learns proposition A, we can determine the entropy of decision problem Q *conditional* on learning A, $E_A(Q)$, as follows:

$$E_A(Q) \quad = \quad \sum_{q \in Q} P(q/A) \times -log_2 P(q/A)$$

In terms of this notion we can now define what might be called the *Relevance* of proposition A with respect to partition Q, $R_Q(A)$, as the *reduction of entropy*, or uncertainty, of Q when A is learned:[3]

$$R_Q(A) \quad = \quad E(Q) - E_A(Q)$$

Relevance will be used to determine the actual interpretation of a sentence underspecified by its conventional meaning. We will say that interpretation A is better than interpretation B, $A > B$, iff $R_Q(A) > R_Q(B)$ with respect to all probability functions for which Q has maximal entropy.

It might be, of course, that for some probability distributions A is better, while for others B is. Which one is then preferred? In those cases, I propose, interpretation A is better if the sentence 'gives rise' to a new question, Q', which is orthogonal to Q, such that after learning A, but not after learning B, every complete answer to Q' also completely answers Q. This indirect notion of relevance will be crucial to account for the implicatures of disjunctive and conditional sentences.

2.7 Why Exhaustify

Consider question (8):

(8) Whom of John and Bill are sick?

This question gives rise to a partition with 4 cells. Assuming that the probability that John is sick equals the probability that Bill is sick, but that the sickness of the one is independent of the other, it is easy to see that the entropy of the question is 2: the question implicitly asks for answers to *two* independent binary questions. Notice that after learning that *(At least) John is sick* the entropy of the question reduces to 1, which means that the relevance of this answer is 2 - 1 = 1. After learning of each of John and Bill whether they are sick, however, the question/decision problem is resolved: the entropy reduces to 0, and the reduction of entropy, the *relevance* of an answer like *John and Bill are sick*, is 2 - 0 = 2. Thus, for an answer to have maximal relevance, it should say of each individual in the domain of quantification whether that individual is sick or not. It should be obvious that this means that complete, or exhaustive, answers to questions are always at least as relevant as partial answers.

Now consider answers (9a) and (9b) to question (8)

(9) a. John is sick.

b. A man is sick.

What is the relevance of these answers, i.e., in how far do these answers reduce the entropy of the question? That depends on how we interpret them. If we interpret them non-exhaustively, the conditional entropy of (8) given (9a) is $(P(J \wedge B/J) \times -log_2 P(J \wedge B/J)) + (P(J \wedge \neg B/J) \times -log_2 P(J \wedge \neg B/J)) = ((\frac{1}{2} \times -log_2 \frac{1}{2}) + (\frac{1}{2} \times -log_2 \frac{1}{2})) = -Log_2 \frac{1}{2} = 1$. Similarly, given that John and Bill are the only men, the conditional entropy of (8) given (9b) is $(P(J \wedge B/J \vee B) \times -log_2 P(J \wedge B/J \vee B)) + (P(J \wedge \neg B/J \vee B) \times -log_2(P(J \wedge \neg B/J \vee B)) + (P(\neg J \wedge B/J \vee B) \times -log_2(P(\neg J \wedge B/J \vee B)) = 3 \times (\frac{1}{3} \times -log_2 \frac{1}{3}) = -log_2 \frac{1}{3} < 1$. The relevance of these two answers according to their non-exhaustive interpretation are thus 1 and something less than 1, respectively. What if we interpret the answers exhaustively? That is, what is the reduction of entropy if we assume that the propositions expressed by the answers are determined after we have applied the exhaustivity operator to (9a) and (9b), respectively? After exhaustification, answer (9a) really means *John is sick and Bill is not*, and after this information is received the entropy reduces from 2 to 0; its relevance is thus 2. Similarly, answer (9b) really means that either only John is sick or that only Bill is sick, and this new information reduces the entropy of the original question from 2 to 1. The important fact to note here is that in both cases the reduction of entropy of the answer under its exhaustive interpretation is *higher* than the reduction of entropy under its non-exhaustive interpretation. And this is in general the case: most answers have a *higher relevance* on their exhaustive reading than on their non-exhaustive reading. On the assumption that speakers are relevance maximizers this means that in case answerers are expected to be cooperative we should interpret these answers exhaustively.

For disjunctive and conditional sentences we have to look at our indirect method. If the question is whether A is the case, A?, and the answer *Yes, or B*, it might be the case that the entropy decreases more on the inclusive reading than on the exclusive reading. Something similar happens with respect to the conditional and biconditional interpretations of answer *If B*. It is natural to assume, however, that both questions 'give rise' to another question: B?. Only on the exclusive and biconditional interpretation of the two answers, every answer to the second question will also resolve the original question whether A is the case. For this reason, the exclusive and biconditional interpretations are preferred.

2.8 Why not always exhaustify

Above, we have criticized Groenendijk & Stokhof's (1984) assumption that in case answers are not explicitly marked as being partial answers, we should always read them exhaustively. One complaint was that this assumption is just an *ad hoc* stipulation. Groenendijk & Stokhof agree, and explicitly regret that they see no way to *derive* exhaustification from the Gricean maxims of conversation, in particular not from Grice's maxim of *quantity.* This complaint can now be met: I have shown in this section that we can motivate the assumption that answers should be read exhaustively by deriving it from the much more general assumption that speakers are relevance optimizers.

What about the other complaint mentioned earlier? As noted above, an answer like *Around the corner* intuitively resolves question *Where can I buy an Italian newspaper?* although it does not suggest that you can buy an Italian newspaper around the corner only. There exists a closely related, and perhaps more obvious, problem for our account suggested so far. Suppose that John asks (10a) and Bill answers by saying (10b) this doesn't (necessarily) mean that we can conclude that (10c) is true.

(10) a. What did Mary wear?

b. A beautiful dress.

c. Mary didn't wear shoes.

This is predicted, however, by the analysis presented so far. The way to account for such problems, or so I argued in van Rooy (to appear), is to realize that questions by themselves are also asked to resolve a decision problem. In case only truth is at issue, such a decision problem can be modeled by a question as well. The usefulness of a question Q' with respect to another question Q, $U_Q(Q')$, can then be determined as the *expected*, or *average*, relevance of the answers to Q' with respect to question Q:[4]

$$U_Q(Q') \quad = \quad \sum_{q' \in Q'} P(q') \times R_Q(q')$$

On the assumption that the actual interpretation of an interrogative sentence is underspecified by its conventional meaning, I have argued that the actual interpretation is the one with the highest relevance. To determine the *relevance* of a question I have argued that for reasons of economy a question should, if possible, not ask for useless information. To implement this I make use of the standard inclusion relation between questions:

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

Now we can count question Q' at least as relevant with respect to question, or decision problem, Q as question Q'' if it holds that:

$$R_Q(Q') \geq R_Q(Q'') \quad \text{iff} \quad \begin{array}{ll} \text{(i)} & U_Q(Q') \geq U_Q(Q''), \text{ and} \\ \text{(ii)} & \text{if } U_Q(Q') = U_Q(Q''), \text{ then } Q' \sqsupseteq Q'' \end{array}$$

The most obvious way that the interpretation of an interrogative is underspecified by its meaning is by the choice of domain over which the *wh*-phrase ranges. If John's decision problem, or question in the background, for (10a) only cares about dresses and trousers, the domain over which the *wh*-phrase *what* ranges should not contain any socks and shoes. Otherwise the interpretation of the question would be less relevant than when these tokens were not taken into account. The exhaustivity operator is now sensitive to this limited set of objects only. But once that is so, we cannot conclude from the exhaustive interpretation of Bill's answer (10b) to John's question (10a) that (10c) is true, just like desired. Thus, although (10b) should be interpreted exhaustively, the exhaustive interpretation involves objects *relevant* to the decision problem only.

Also the problematic mention-some phenomena can be accounted for when we assume that speakers are relevance optimizers. For these examples, however, it seems that exhaustification doesn't really seem to play a role. In van Rooy (to appear) I argue that mention-some questions are asked, or mention-some answers are given, only in particular circumstances, and show that in these circumstances the utility of mention-some questions/answers coincide with their mention-all alternatives. For reasons of economy, mention-some readings are in these circumstances preferred. We can conclude that although in normal circumstances (or better, perhaps, out of context) the exhaustive reading of an answer is more relevant than its non-exhaustive counterpart, in special circumstances it is not. As a result, we can explain that for reasons of optimizing relevance, exhaustification does not always take place, and even when it takes place, it cares about *relevant* objects only.[5]

3. Generalized Conversational Implicatures

3.1 Horn's division of labor

Consider a typical case of communication where two meanings m_i and m_j can be expressed by two linguistic forms f_i and f_j. In principle this gives rise to two possible codings: $\{\langle f_i, m_i\rangle, \langle f_j, m_j\rangle\}$ and $\{\langle f_i, m_j\rangle, \langle f_j, m_i\rangle\}$. In many communicative situations, however, the underspecification does not really exist, and is resolved due to the general pragmatic principle that a lighter form will be interpreted by a more salient, or stereotypical, meaning:[6] (i) It is a general defeasible principle, for instance, in centering theory that if a certain object/expression is referred to be a pronoun, another more salient object/expression should be referred to by a pronoun too; (ii) Reinhard (1983) and Levinson (1987) seek to reduce Chomsky's B and C principles of the binding theory to pragmatics maxims. In particular, disjoint reference of lexical NPs throughout the sentence is explained by pointing to the possibility of the use of a lighter expression, viz. an anaphor or pronoun; (iii) The preference for bridging (Clark & Haviland, 1977) and stereotypical interpretations (Atlas & Levinson, 1981); (iv) and perhaps most obviously, Horn's (1984) division of pragmatic labor according to which a marked expression (morphologically complex and less lexicalized) typically gets a marked meaning (cf. *John made the car stop* versus *John stopped the car*). In neo-Gricean pragmatics proposed by Atlas, Horn and Levinson, this principle is explained through the interaction of the so-called Q and I principles, and has recently been incorporated in (bi-directional) optimality theory by Blutner (2000) and reformulated in terms of game theory by Dekker & van Rooy (2000). However, as we have seen above, explanations based on the Q and I principles are very shaky: these principles tend to clash with one another, and it is not always clear how to resolve this clash. In particular, it's unclear under which circumstances which principle should be used to explain the phenomena. I will show that by thinking of language as an efficient coding system the principle that lighter expressions get a more salient meaning can be given a straightforward explanation. But I will also relate this communication-theoretic efficient coding analysis of Horn's division of pragmatic labor with a recent Game-Theoretical analysis proposed by Prashant Parikh.

3.2 Parikh's Game-theoretical analysis

3.2.1 Horn's division and Optimal expected utility.

Parikh (2000) wants to give a game-theoretical underpinning of why an in principle ambiguous, or underspecified, sentence typically gets interpreted in its most likely way. The argument goes roughly as follows: A speaker used an expression, f, that in principle could be interpreted in several ways. How f in fact should be be interpreted depends on the actual situation the speaker is in. This actual situation can be thought of either as the situation *external* to the agent, or as some *internal* state of the agent. We will interpret things in the latter way, and think of the situation as the intention of the agent to express a certain meaning: m or m'. The hearer doesn't know, of course, which meaning the speaker wants to express, but thinks that expressing m is more likely than expressing m'. For concreteness, let us assume that the probabilities he assigns to m and m' are 0.8 and 0.2, respectively, and that this is common knowledge. Parikh assumes that for interpreting an expression, we also have to take into account the *alternative expressions* that the speaker might have used. It is assumed that besides the underspecified form f, there are also expressions f' and f'' that can each have one meaning only: f' can only mean m' and f'' only m. Parikh proposes that whether f should be interpreted as m or not depends on whether the form-meaning pair $\langle f, m\rangle$ is part of the optimal combined speaker and hearer 'strategy' to resolve the communication game associated with the situation described above.

In the communicative game the speaker is in a certain internal state, he wants to express m or m', and wants to communicate this to the hearer. He has to choose between three forms: f, f' or f''. After he made his utterance, the hearer has to interpret the form either as m or m'. She doesn't know which state the speaker is in. Although the action chosen by the hearer might depend on the action of the speaker, we might model the game as one in which they make their choices *simultaneously*. To do so, however, we have to assume that they choose **strategies** rather than concrete actions. A strategy consists of a rule that determines what a player will do in different circumstances. A speaker's strategy, S, is a function from meanings to forms, i.e. an element of $[\{m, m'\} \rightarrow \{f, f', f''\}]$, and a hearer's strategy, H, is a function from forms to meanings, i.e. an element of $[\{f, f', f''\} \rightarrow \{m, m'\}]$. In a table, this can be displayed as follows:

Speaker :

	m	m'
S_1	f	f'
S_2	f	f
S_3	f''	f
S_4	f''	f'

Hearer :

	f	f'	f''
H_1	m	m'	m
H_2	m'	m'	m

Which meanings should be associated with which forms depends on which speaker-hearer strategy pair is the *solution* of this game. Such a solution should be a (Nash) *equilibrium*. The search for equilibria involves the search for an optimal combination of a speaker strategy and a hearer strategy: $\langle S, H\rangle$ is a (Nash) equilibrium if and only if:

$$\begin{array}{ll} \text{(i)} & \neg\exists S' : \langle S', H\rangle > \langle S, H\rangle \\ \text{(ii)} & \neg\exists H' : \langle S, H'\rangle > \langle S, H\rangle \end{array}$$

Notice that the search for an equilibrium requires an ordering relation between speaker-hearer strategy pairs. Parikh proposes to define this ordering in terms of the *expected utilities* of the speaker-hearer strategies:

$$EU(S, H) \quad = \quad \sum_m P(m) \times U(m, S, H)$$

Notice that in this definition I have followed Parikh assuming that only one probability function and only one utility function are relevant. The *probability* function is the one from the *hearer*, and speaker and hearer are assumed to have the *same utility function* because communication is a game of coordination.

Parikh proposes that the utility function is sensitive to the *complexity* of the expressions involved in the following way: successful communication is most important, but success with a simple expression (by using f) is preferred to success with a complex expression (by using f' or f''). Let us assume that the complexity of a form can be measured by a natural number and that $Compl(f) = 1$, while $Compl(f') = Compl(f'') = 2$. Making use of this complexity function, Parikh's utility function can be defined as follows:

$$\begin{array}{lcl} U(m,S,H) & = & Compl(S(m))^{-1}, \text{ if } H(S(m)) = m \\ & = & 0 \text{ otherwise} \end{array}$$

Now we can calculate for all the combined speaker-hearer strategies their expected utilities as lotteries over the utilities of these strategies in states m and m', and see which of them form Nash equilibria:

m:

m	H_1	H_2
S_1	1	0
S_2	1	0
S_3	0.5	0.5
S_4	0.5	0.5

m':

m'	H_1	H_2
S_1	0.5	0.5
S_2	0	1
S_3	0	1
S_4	0.5	0.5

EU:

	H_1	H_2
S_1	**0.9**	0.1
S_2	0.8	0.2
S_3	0.4	**0.6**
S_4	0.1	0.5

The game has two Nash equilibria: $\langle S_1, H_1\rangle$ and $\langle S_3, H_2\rangle$. Notice that according to the first one, $\langle S_1, H_1\rangle$, the more probable state, or meaning, m, is expressed by the lighter form f, while the less probable state, or meaning, m', is expressed by the complex form f'. According to the other Nash equilibrium, however, the more probable meaning is expressed by a more complex form, while the less probable meaning is expressed by a lighter form. Thus, only if speaker and hearer coordinate on the first Nash equilibrium, we can give a game-theoretical explanation of Horn's division of pragmatic labor. Parikh proposes to explain this by assuming that conversational participants take only that Nash equilibrium into account which has the *highest expected utility*: $\langle S_1, H_1\rangle$.

3.2.2 Generalizing Horn's division. In the previous section we have seen how we can account for the intuition that more probable meanings get expressed by lighter forms. That one meaning is more probable than another typically means that this meaning is more salient or stereotypical than the other, and that was what Horn's division was about. But Parikh's approach does not only make interesting predictions in case we vary the probabilities: a kind of division of labor occurs already just because of the utilities involved.

In the first part of this paper we have seen that in the same situation different interpretations of a single sentence might be possible; that these different interpretations might all have a positive value; but that one of those interpretations is still better – more relevant – than another. To formalize this in Parikh's framework, we have to make a distinction between *states* and *meanings*: different meanings can have a positive utility in the same state. Assuming that we have 2 states, t and t', 2 meanings, and 3 forms, the speaker and hearer strategies are as follows:

Speaker :

	t	t'
S_1	f	f'
S_2	f	f
S_3	f''	f
S_4	f''	f'

Hearer :

	f	f'	f''
H_1	m	m'	m
H_2	m'	m'	m

To let utility and costs do the work, we will assume that t and t' are equally likely: $P(t) = P(t') = 0.5$. Just for concreteness, I will assume that the utilities of the 2 meanings in the two states are as follows: $U(m,t) = 10$, $U(m,t') = 10$, $U(m',t) = 4$, $U(m',t') = 10$. You might think of m as the mention-all reading of a question, or the exhaustive answer to a question, and m' as its mention-some reading, or partial answer. The utility distribution reflects the fact, for instance, that the utility of a mention-some/non-exhaustive reading might be equally high as its corresponding mention-all/exhaustive reading (in t'), but can be lower too (in t). But utility is not all what counts here: we have to take into account the *costs* as well. Let us assume for concreteness that the hearer's cost for interpreting (or answering) a sentence as m is 3, $c_h(m) = 3$, while interpreting it as m' is 0, $c_h(m') = 0$. Again, this seems reasonable: interpreting a question/assertion exhaustively takes more effort than a non-exhaustive interpretation, because on the former reading more individuals have to be taken into account. For the *speaker* it are the costs of uttering the form that counts. Let us assume that $c_s(f) = 0$ and $c_s(f') = c_s(f'') = 2$. This is a similar assumption as Parikh (2000) makes, so we don't have to motivate this anymore. With the probabilities, utilities, and costs set in this way, we can determine the utilities of the speaker-hearer strategies in the different states and their expected utilities, if we assume that for expected utility the cost of the speaker and the hearer should be counted up:

t:

t	H_1	H_2
S_1	7	4
S_2	7	4
S_3	5	5
S_4	5	5

t':

t'	H_1	H_2
S_1	8	8
S_2	7	10
S_3	7	7
S_4	8	8

EU:

	H_1	H_2
S_1	7.5	6
S_2	7	6
S_3	6	6
S_4	6.5	6.5

This game has two Nash-equilibria, but the one with the highest expected utility is the one we were after: $\langle S_1, H_1 \rangle$. According to this combined strategy, the meaning which has the highest (expected) utility, or relevance, is expressed by the lighter expression. More importantly, I think, is the other side of the story: if you want to express a meaning that is *less relevant*, you have to use a *marked form*.[7] This prediction is, arguably, born out by the facts: When you ask a *wh*-question out of context you typically have to mark the fact that you are not interested in its exhaustive answer:

(11) Who, *for example*, came to the party?

Groenendijk & Stokhof (1984) have noted, similarly, that to ask a mention-some question in Dutch, you typically don't use a *wh*-interrogative. To ask the mention-some reading of question (12a) in Dutch, you should not ask the *wh*-question (12b), but rather a *yes/no*-question like (12c):

(12) a. Who has a light?

b. Wie heeft er een vuurtje voor me?

c. Heeft er iemand een vuurtje voor me?

Something similar holds, arguably, for answers: In the context of a single question an answer should normally be interpreted exhaustively, i.e. in the most relevant way, and a non-exhaustive (or non-complete) answer/reading has to be marked. Incomplete answers are typically fronted by the interjector *well*,[8] and Büring (1999), for instance, has argued that partial answers are marked by *rising*, instead of falling, intonation of the relevant terms. Even more speculative: it is generally assumed that *topic change* has to be marked in a conversation. That seems to fit our picture too, because when a sentence changes the topic it is obviously less relevant with respect to the earlier topic than when it doesn't.

3.2.3 Conversational versus Conventional. Although the extension of Horn's division of pragmatic labor to relevance-phenomena seems appealing, the use of Parikh's framework to account for it is, on second thought, somewhat problematic. For the analysis it was crucial that speaker's cost of uttering and hearer's costs of interpreting had to be counted up to determine the expected utilities. For the speaker and hearer themselves, however, there seems no reason why this should be done for a *particular* conversational situation they might be in. The crucial supposition seems more reasonable, though, when expected utilities are determined on the assumption that agent's play both speaker-and hearer-roles, and do so both half of the time. When we make that assumption, however, we are outside the domain of *particularized* conversational implicatures, and add a *conventional* feature.[9]

In fact, I believe that Parikh's explanation of Horn's division of labor given in section 3.2.1 suggests already that the implicatures based on Horn's division depend less on the actual conversational situation than Parikh seems to assume. The reason is that Parikh (2000) assumes that the payoffs of speaker and hearer are – or are nearly – the same, and that for $\langle S, H\rangle$ to be the solution of the game, this strategy pair has to have the highest expected utility. In the next section I will show that once these assumptions are made, plus the assumption that there are 'enough' forms to express meanings, Parikh's analysis comes down to a standard Communication Theoretic analysis of *optimal coding* of information, which should be classified as *conventional* rather than conversational in nature.

3.3 Optimal coding of information

The games played by Parikh reviewed in section 3.2.1 always have two Nash equilibria: one where m is expressed by the underspecified form f and m' by the more complex f', and one where m' is expressed by f and m' by f''. This analysis is based on the assumption that f' and f'' have already a fixed meaning. But once we assume that we are after a one-to-one correspondence between meanings and forms, we don't need to assume that there are 3 forms involved such that the 2 meanings each have already at least one form with that fixed meaning: we might as well start with 2 forms only whose meanings are *both* underspecified by compositional semantics, and determine the actual interpretations based on the orderings of meanings and forms alone. But this is exactly what optimal coding is about.

The question that started communication theory was: how can we send messages over a channel as quickly as possible without distortion? The answer is: by looking for the optimal coding; to represent the data (or communicate meanings) in a way as comprehensive as possible. Suppose we have a source (a speaker) that sends meanings from a set $M = \{m_1, ..., m_n\}$ in terms of codes built up from code symbols belonging to the code-alphabet $A = \{a_1, ..., a_n\}$. A *source code*, or *coding system* is defined as a function from M to $F = A^*$, where '$*$' is Kleene's star. Thus, a coding system can be equated with a speaker's strategy, S – or better a combined speaker-hearer strategy according to which different meanings are expressed by different forms – in Parikh's game. For example, $S(\text{red}) = 00$, $S(\text{white}) = 11$, $S(\text{blue}) = 10$, $S(\text{orange}) = 01$ is a source code for $M = \{\text{red, white, blue, orange}\}$ with alphabet $A = \{0, 1\}$. Of course, the sets M and $F = A^*$ allow for many different codes/speaker's strategies. Intuitively, however, some codings are more *efficient* than others. What is the best strategy? The one that gives rise to codes with the shortest *expected length.* The crucial insight of Shannon (1948) was that this expected length depends not only on the *length* of the messages after encoding, but also on the *probability* with which the messages are sent. Suppose that P is a probability distribution over M. Suppose, moreover, that $l(S(m))$ is the length of the codeword associated with m. In that case, the *expected*, or average, *length* of a source code S (or average length of speaker-hearer strategy pair $\langle S, H\rangle$) for M and P, $EL(S)$, is given by:

$$EL(S) \quad = \quad \sum_{m \in M} P(m) \times l(S(m))$$

To illustrate, let us extend our example by assuming that the probability distribution of M is $P(\text{red}) = \frac{1}{2}$, $P(\text{white}) = \frac{1}{4}$, $P(\text{blue}) = \frac{1}{8}$, $P(\text{orange}) = \frac{1}{8}$. Then we can easily see that the coding $S'(\text{red}) = 0$, $S'(\text{white}) = 10$, $S'(\text{blue}) = 110$, $S'(\text{orange}) = 111$ has a shorter expected length than the coding given above: 1.75 to 2. A crucial difference between the two coding strategies is that in distinction with S, S' does not encode all elements of M with the same length: the more probable elements of M get an encoding with a shorter length.[10]

Now it is easy to show that the assumption of optimal coding accounts for Horn's observation that simple expressions get a salient or stereotypical interpretation, while complex expressions a marked one. We just have to replace *length* by *complexity.* Suppose that the conventional meanings of forms f_i and f_j are such that they both could express m_i and m_j. For instance, with both words *kill* and *cause to die* we could denote situations of *direct* (stereotypical) and *indirect* (marked) killing, and with both unstressed *he* and stressed *HE* we could refer to both *salient* and *non-salient* male individuals in the discourse. Still, the less complex *kill* will typically be interpreted as direct stereotypical killing, and the other way around for complex *cause to die.* And this follows from the assumption that speakers use a language that optimally encodes the relevant information. In this case we have two relevantly different coding systems: S, which assigns m_i to f_i and m_j to f_j, and S', which assigns m_i to f_j and m_j to f_i. The probabilities are such that $P(m_i) > P(m_j)$ and – denoting the complexity of f by $C(f)$ – we also assume that $C(f_i) < C(f_j)$. A standard proof showing that $EC(S') - EC(S) > 0$ demonstrates then that S is a more optimal coding than S':

$$\begin{aligned} EC(S') - EC(S) &= \sum_m P(m) \times C(S'(m)) - \sum_m P(m) \times C(S(m)) \\ &= (P(m_i) \times C(f_j)) + (P(m_j) \times C(f_i)) \\ &\quad -(P(m_i) \times C(f_i)) - (P(m_j) \times C(f_j)) \\ &= (P(m_i) - P(m_j)) \times (C(f_j) - C(f_i)) \end{aligned}$$

Because $P(m_i) - P(m_j) > 0$, S' can only be more optimal than S in case $C(f_j) - C(f_i) < 0$. But this is by assumption not the case, and so S is preferred to S'. The same proof shows that when $P(m_i) - P(m_j) > 0$ the optimal code is such that $C(f_i) \leq C(f_j)$; only then $EC(S') - EC(S) \geq 0$.[11] Thus, the meanings that are more likely to be communicated will be encoded with a smaller complexity.

Notice that by the way Parikh (2000) defined his utility function, it will hold that for every $m \in M$ it is the case that whenever $C(S(m))$ *in*creases, $U(m, S, R)$ will *de*crease. Thus, $C(S'(m) - C(S(m)) > 0$ if and only if $U(m, S, R) > U(m, S', R') > 0$. From this we can conclude that whenever the *expected complexity* of the forms according to a strategy will increase - i.e. whenever $EC(S)$ gets higher – also the *expected utility* of the corresponding speaker-hearer strategy, $EU(S, R)$, will decrease.[12] But this means that – if we consider only one-to-one mappings between meanings and forms – Parikh's analysis in terms of *maximization* of expected utility is essentially the same as an optimal coding analysis in terms of *minimization* of expected complexity.

Notes

1. But note that van Kuppevelt doesn't make use of Groenendijk & Stokhof's exhaustification operator, and in contrast to them also claims that exhaustive interpretation has truth-conditional impact.

2. See van Kuppevelt (1996), followed by Scharten (1997) and Carston (1998).

3. This notion was used by Lindley (1956) already to measure the informational value of a particular result of an experiment.

4. This measure is standardly known in communication theory as the *mutual information* or *rate of actual transmission* between Q and Q' (cf. Cover & Thomas, 1991).

5. In this paper I have used relevance, or communication theory, to explain *why* exhaustification is such a natural operator to accoount for scalar implicature, and to determine *when* it should be applied. In newer work I rather *define* the exhaustification operator in terms of a notion of relevance. In this way we can improve considerably on Groenendijk & Stokhof's (1984) notion of exhaustivity and can *always* apply the exhaustivity operator to an answer.

6. See Horn (1984) and Levinson (2000) for general discussions.

7. To be honest, this outcome depends crucially on how the utilities and costs are set. I believe, however, that these are natural for the particular examples that we have in mind.

8. The conversational analysist Pomerantz (1984), for instance, argues that *well* signals moves that are in some way or other dispreferred.

9. In van Rooy (to appear 2) I give an *evolutionary* account of Horn's division of pragmatic labor based on the assumption that players play both speaker- and hearer-roles.

10. There exists a close connection with entropy here: for coding S', but not for S, the expected length, $EL(S')$ is equal to the entropy of M. Shannon (1948) showed that this is

the optimal one can reach for w.r.t. uniquely decodable codes. Notice that this means that the optimal codelength for each m equals $-log_2 P(m)$, the surprise value of m. This means that, at least for the optimal coding, there is a direct correspondence between complexity of expression and the informativity of the meaning it expresses.

11. This holds in general: in case $P(m_i) > P(m_j) > P(m_k)$ the optimal coding strategy S will be such that $C(S(m_i)) \leq C(S(m_j)) \leq C(S(m_k))$ (cf. Cover & Thomas, 1991).

12. Here is the straightforward proof:

$$
\begin{array}{lll}
EC(S') > EC(S) & \text{iff} & \sum_m P(m) \times C(S'(m) > \sum_m P(m) \times C(S(m)) \\
 & \text{iff} & \sum_m P(m) \times (C(S'(m)) - C(S(m))) > 0 \\
 & \text{iff} & \sum_m P(m) \times (U(m, S, R) - U(m, S', R')) > 0 \\
 & \text{iff} & \sum_m P(m) \times U(m, S, R) > \sum_m P(m) \times U(m, S', R') \\
 & \text{iff} & EU(S, R) > EU(S', R')
\end{array}
$$

References

Anscombre J.C. and O. Ducrot (1983), *L'Argumentation dans la langue*, Brussels, Mardaga.

Atlas, J. and S. Levinson (1981), 'It-Clefts, Informativeness and Logical Form', In: P. Cole (ed.), *Radical Pragmatics*, New York, AP.

Blutner, R. (2000), 'Some aspects of Optimality in Natural Language Interpretation', *Journal of Semantics*, **17**: 189-216.

Büring, D. (1999), *The meaning of Topic and Focus – The 59th Street Bridge Accent*, London: Rootledge.

Carston, R. (1998), 'Informativeness, Relevance and Scalar Implicature', In: R. Carston & S. Uchida (eds.), *Relevance Theory: Applications and Implications*, John Benjamins, Amsterdam, 179-236.

Clark H. H. & J. Haviland (1977), 'Comprehension and the given-new contract', In R. Freedle (ed.), *Discourse production and comprehension*, Hillsdale, NJ: Lawrence Erlbaum, pp. 1-40.

Cover, T.M. & J.A. Thomas (1991), *Elements of Information Theory*, Wiley: New York.

Dekker, P. & R. van Rooy (2000), 'Bidirectional Optimality Theory: an application of Game Theory', *Journal of Semantics*, **17**: 217-242.

Gazdar, G. (1979), *Pragmatics*, London: Academic Press.

Groenendijk, J. and M. Stokhof (1984), *Studies in the Semantics of Questions and the Pragmatics of Answers*, Ph.D. thesis, University of Amsterdam.

Grice, H. P. (1975), 'Logic and Conversation', In: P. Cole & Morgan (eds.), *Syntax and Semantics 3: Speech Acts*, New York: Academic Press.

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

Kuppevelt, J. van (1996), 'Inferring from Topics: Scalar Implicature as Topic-Dependent Inferences', *Linguistics and Philosophy*, 19, pp. 555-598.

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

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, 11-42, Washington; Georgetown University Press.

Horn, L. (2000), 'From *if* to *iff*: Conditional perfection as pragmatic strengthening', *Journal of Pragmatics*, **32**: 289-326.

Levinson, S.C. (1987), 'Pragmatics and the grammar of anaphora', *Journal of Linguistics*, **23**: 379-434.

Levinson, S.C. (2000), *Presumptive Meanings. The Theory of Generalized Conversational Implicatures*, MIT Press: Cambridge, Massachusetts.

Lindley, D. V. (1956), 'On a measure of information provided by an experiment', *Ann. Math. Stat.*, 29, pp. 986-1005.

Matsumota, Y. (1995), 'The conversational condition on Horn scales', *Linguistics and Philosophy*, **18**: 21- 60.

McCawley, J. (1993), *Everything that Linguists always wanted to know about Logic, but were afraid to ask*, Chicago: Chicago University Press.

Merin, A. (1997), 'Information, relevance, and social decisionmaking', In: L. Moss, J. Ginzburg, M. de Rijke (eds.), *Logic, Language, and Computation, Vol. 2*, Stanford.

Parikh, P. (2000), 'Communication, meaning, and interpretation', *Linguistics and Philosophy*, **23**: 185-212.

Pomerantz, A. (1984), 'Agreeing and disagreeing with assessments: some features of preferred/dispreferred turn shapes', In. J. Atkinson & J. Heritage (eds.), *Structures of social action: Studies in conversation analysis*, pp. 57-101, Cambridge: Cambridge University Press.

Reinhard, T. (1983), *Anaphora and semantic interpretation*, London: Croom Helm.

Rooy, R. van (2001), 'Relevance of communicative acts', In *Theoretical Aspects of Rationality and Knowledge; Proceedings of TARK 2001*, J. van Benthem (ed.), San Francisco, Morgan Kaufmann Publishers, Inc., pp. 83-96.

Rooy, R. van (to appear), 'Utility of mention-some questions', *Language and Computation*.

Rooy, R. van (to appear2), 'Signalling games select Horn strategies', to appear in *Linguistics and Philosophy*.

Scharten, R. (1997), *Exhaustive Interpretation: A Discourse Semantic Account*, Ph.D. thesis, University of Nijmegen.

Shannon, C. (1948), 'The Mathematical Theory of Communication', *Bell System Technical Journal*, 27.