Conversational Implicatures

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 efficient 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 underspecification 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 The Q and I principle

Neo-Gricean pragmatics 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,

[*] This research has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences.

and it allows us to interpret a conditional like 'John walks, if Mary walks' as the biconditional 'John walks if and only Mary walks'.

3 Problems for the Q and I principles

3.1 Too general

Although the Q and I principles are intuitively appealing, they give rise to a number of conceptual and empirical problems. 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'. Horn, Levinson and others have argued that these problems can be prevented by (i) 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. 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.

3.2 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 $\langle$and, or$\rangle$ 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.

4 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.

5 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.

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

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 $\langle$all, some$\rangle$ scale. In contrast to Gazdar's analysis of scalar implicatures, however, our analysis does not give rise to the wrong prediction that John is not coming: the exhaustive reading of *A man is coming* as answer to question *Who comes?* is compatible with the fact that John is. 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 beans, 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 that the property underlying a question like (1a) is $\lambda w.Walk(j)(w)$, and that answer (1b) should be represented by $\lambda P.\forall w[Talk(m)(w) \rightarrow P(w)]$ which after exhaustification means that Mary talked iff John walked.[1]

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. When 3 flags are mutually known by us to be all white except for a small block of some other distinguishing color (being either red, yellow or green), and 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*, while we don't even generate the implicature because we can assume that the implicit question was something like *What is the color of the small block?*

Indeed, our topic-dependent analysis of 'scalar' implicatures prevents us from triggering implicatures to be *cancelled* later for reasons of relevance (see also van Kuppevelt (1997) and Carstyn (ms)). Consider the following 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. If we allow for explicit quantification over worlds, we can represent an answer like A *or* B *or* C in terms of an existential quantifier as follows: $\lambda P.\exists w[(A(w) \vee B(w) \vee C(w)) \wedge P(w)]$. If

[1] This inference is accounted for by Groenendijk & Stokhof (1984) in terms of their generalized exhaustivity operator without using explicit quantification over worlds. Such an analysis cannot account, however, for the exclusive reading of disjunctive sentences with more than two disjuncts, to be discussed below.

we now assume that this sentence is given as answer to the question 'What proposition(s) is/are true?', exhaustivity has the effect that only worlds count that make just one of the three propositions true, resulting 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 in (4), where the complex sentence is given as an (exhaustive) polar answer:

(4) 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:

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

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

6 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 exhaustivily 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 (6a) and you answer by saying (6b), 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.

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

b. Around the corner.

6.1 Topic dependent relevance

In cooperative dialogues the relevance of communicative acts can be determined with respect to *decision problems* (cf. van Rooy (2001). A decision problem Using 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:

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

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:[2]

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

Relevance will be used to determine the actual interpretation of a sentence underspecified by its conventional meaning. We will say

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

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.

6.2 Why Exhaustify

Consider question (7):

(7) 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 (8a) and (8b) to question (7)

(8) 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 (7) given (8a) 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 (7) given (8b) 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 (8a) and (8b), respectively? After exhaustification, answer (8a) 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 (8b) 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.

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 I 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. Fortunately, however, also the problematic mention-some phenomena can be accounted for when we assume that speakers are relevance optimizers. In van Rooy (to appear) I argue that mention-some questions are asked, or mention-some answers are given, only in very particular circumstances, and show that in these circumstances the utility, or relevance, 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 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.

7 Explaining markedness

7.1 Horn's division of labor

Consider a typical case of communication where two meanings m_1 and m_2 can be expressed by two linguistic representations r_1 and r_2. In principle this gives rise to two possible codings: $\{\langle r_1, m_1\rangle, \langle r_2, m_2\rangle\}$ and $\{\langle r_1, m_2\rangle, \langle r_2, m_1\rangle\}$. 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: (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 marked expression (morphologically complex and less lexicalized) typically get a marked meaning (cf. *John made the car stop* versus *John stopped the car* and consider also the fact that stressed pronouns can pick up less salient objects). 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.

7.2 Optimal coding of information

The question that started information 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 in a way as comprehensive as possible. Suppose we have a source (without memory) that sends messages from a set $U = \{u_1, ..., u_n\}$ in terms of codes built up from codesymbols belonging to the code-alphabet $S = \{s_1, ..., s_n\}$. A *source code*, or *coding system*, C, is defined as a function from U to S^*, where '$*$' is Kleene's star. For example, $C(\text{red}) = 00$, $C(\text{white}) = 11$, $C(\text{blue}) = 10$, $C(\text{orange}) = 01$ is a source code for $U = \{\text{red, white, blue, orange}\}$ with alphabet $S = \{0, 1\}$. Of course, the source and alphabet allow for many different codes. Intuitively, however, some codings are more *efficient* than others. What is the best coding system? The coding system 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 function P assigns numbers to the elements of U such that $\sum_{u \in U} P(u) = 1$, i.e. suppose that P is a probability distribution over U. Suppose, moreover, that $l(C(u))$ is the length of the codeword associated with u. In that case, the *expected length* of a source code C for U and P is given by:

$$L(C) \quad = \quad \sum_{u \in U} P(u) \times l(C(u))$$

To illustrate, let us extend our example by assuming that the probability distribution of U 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 $C'(\text{red}) = 0$, $C'(\text{white}) = 10$, $C'(\text{blue}) = 110$, $C'(\text{orange}) = 111$ has a shorter expected length than the coding given above: 1.75 to 2. A crucial difference between the two codings is that in distinction with C, C' does not encode all elements of U with the same length: the more probable elements of U get an encoding with a shorter length.[3] This holds in general: in case $P(u_i) > P(u_j)$ the optimal coding C will be such that $l(C(u_i)) \leq l(C(u_j))$ (cf. Cover & Thomas, 1991). Thus, the messages that are more likely to be sent will be encoded with a smaller length. This fact can now be used to account for Horn's division of pragmatic labor. Suppose that speakers have the following set of *contents/meanings* that they might want to communicate: $M = \{m_1, ..., m_n\}$. On average, we might assume that the probabilities with which they want to communicate these contents/meanings are correlated with the probabilities with which these contents are true: if m_i is more likely to be the case, or more *stereotypical*, than m_j, the probability that speakers want to communicate m_i is higher than that of m_j. In other communicative situations the probability with which the elements of M are communicated depends on how *salient* the elements of M are.[4] Speakers cannot send the contents without representing them. Let us assume that speakers can use the elements of R (the *representations*), $R = \{r_1, ..., r_k\}$, to encode the elements of M. The elements of R might be varied: some are more *complex* than others. Codings are now functions from M to R. Just as before we can ask: what is the best coding? Following standards in data comprehension, the answer is: the coding that *minimizes average complexity.* Average complexity of coding system C

[3] There exists a close connection with entropy too: for coding C', but not for C, the expected length, $L(C')$ is equal to the entropy of U, $E(U)$. It turns out 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 u_i is equal to $-log_2 P(u_i)$, the surprise value of u_i.

[4] In yet others, the probabilities depend even more on the conversational situation and correlate with relevance.

is defined as follows:

$$Compl(C) \quad = \quad \sum_{m \in M} P(m) \times Compl(C(m))$$

Now it is easy to show that the assumption of optimal coding accounts for Horn's observation that simple expressions get a salient/stereotypical interpretation, while complex expressions a marked one. Suppose that the conventional meanings of representations r_i and r_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, an 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: C, which assigns m_i to r_i and m_j to r_j, and C', which assigns m_i to r_j and m_j to r_i. The probabilities and complexities are such that $P(m_i) > P(m_j)$ and $Compl(r_i) > Comp(r_j)$. A standard proof showing that $Compl(C') - Compl(C) > 0$ demonstrates then that C is a more optimal coding than C' (where I abbreviate $P(m_i)$ by p_i, $Compl(r_i)$ by cpl_i, and '$\times$' by '$\cdot$'):

$Compl(C') - Compl(C) =$
$= \sum p_i \cdot Compl(C'(i)) - \sum p_i \cdot Compl(C(i))$
$= (p_i \cdot cpi_j) + (p_j \cdot cpl_i) - (p_i \cdot cpl_i) - (p_j \cdot cpl_j)$
$= (p_i - p_j) \cdot (cpl_j - cpl_i)$

Because $p_i - p_j > 0$, C' can only be more optimal than C in case $cpl_j - cpl_i < 0$. But this is by assumption not the case, and so C is preferred to C'. The same proof shows that when $p_i - p_j > 0$ the optimal code is such that $cpl_i \leq cpl_j$; only then $Compl(C') - Compl(C) \geq 0$. Horn's division explained.

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.*

Carston, R. (ms.), *Informativeness, Relevance and Scalar Implicature*, University College London.

Clarck 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.*

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.

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.*

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