Relevance Implicatures

Robert van Rooy[*]
ILLC/University of Amsterdam
vanrooy@hum.uva.nl

Abstract

According to standard pragmatics, we should account for conversational implicatures in terms of Grice's (1967) 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) in fact, we can better account for implicatures by means of an exhaustivity operator defined in terms of a *relevance*-based ordering relation.

1 Introduction

In most circumstances we infer from the fact that the (a) sentences of the following examples are uttered that the (b) sentences are true:

(1) a. John ate *some* of the cookies.

b. John didn't eat *all* of the cookies

(2) a. John is coming *or* Mary is going.

b. John is coming or Mary is going, but not *both*.

(3) a. John is walking, *if* Mary is talking.

b. John is walking *if and only if* Mary is talking.

[*] This note reports on work in progress on the use of relevance for the semantic analysis of natural language. Parts of this paper was presented at the second workshop on Optimality Theoretic Semantics at OTS in Utrecht (2000) and at the Formal Pragmatics conference in Berlin (2001). I would like to thank the participants of those events for their comments and suggestions. I would also like to thank the anonymous reviewers, Reinhard Blutner and expecially Katrin Schulz. I am grateful to William Rose for discussion and for correcting my English. The research for this work is supported by a fellowship from the Royal Netherlands Academy of Arts and Sciences (KNAW), which is gratefully acknowledged.

Griceans argue that these inferences should not be accounted for in terms of semantic entailment because the inferences do not always go through. Instead, they argue that the (b) sentences are only *conversationally implicated*, and suggest that this should be accounted for in terms of Grice's maxim of *Quantity.* As a result, they can account for the above inferences without giving up the standard semantic treatment of *some*, *or*, and *if.*

Over the years this intuitive explanation has been formalized and treated as implicatures that (i) are induced by the two Gricean submaxims of *Quantity*; (ii) are *general* and associated with specific lexical items; (iii) can be *cancelled* for reasons of relevance; but (iv) are truth-conditionally irrelevant. In this paper I wish to argue, instead, that the inferences can be explained by an assumption of *optimal relevance*; due to *particular* contextual features and thus *not cancellable* for reasons of relevance; and that the inferences might well be *truth-conditionally relevant.* Still, as we will see, this doesn't mean that we have to give up the standard analysis of *some, or* or *if.*

2 The Q and I principles

2.1 Towards Generalized Q- and I-implicatures

Perhaps the most important notion in linguistic pragmatics is Grice's (1957) notion of *conversational implicature.* It is based on the insight that by means of general principles of rational communication we can communicate more with the *use* of a sentence than the *conventional meaning* associated with it. What is communicated depends not only on syntactic and semantic rules, but also on facts about the utterance situation, the linguistic context, and the goals and preferences of the interlocutors of the conversation.

Grice assumes a theoretical distinction within the 'total significance' of a linguistic utterance between what the speaker *explicitly said* and what he has merely *implicated.* What has been said is supposed to be based purely on the *conventional* meaning of a sentence, and is the subject of compositional semantics. What is implicitly conveyed belongs to the realm of pragmatics. These implicatures are based on Grice's *cooperative principle*: the assumption that speakers are maximally efficient rational cooperative language users. Grice comes up with a list of four rules of thumb – the maxims of *quality*, *quantity*, *relevance*, and *manner* – that specify what participants have to do in order to satisfy this principle. They should speak sincerely, relevantly, and clearly, and should provide sufficient information.

Over the years many phenomena have been explained in terms of the Gricean maxims of conversation. Horn (1972) and especially Gazdar (1979) proposed to formalize Grice's suggestions in order to turn informal pragmatics into a predictive theory. They concentrated on Grice's maxim of quantity, and especially on its first submaxim.

- **Quantity**

  1. Make your contribution as informative as is required

(for the current purposes of the exchange).

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

Atlas & Levinson (1981), Horn (1984) and Blutner (1998) tried to formally account not only for Quantity 1 implicatures, but also for implicatures that appeal to Grice's second Quantity maxim and the maxims of Relation and Manner. They developed an account which maintains Grice's maxim of Quality but replaces all his other maxims with two general principles:

- Q: Make your contribution sufficient; say as much as you can (given I).
- I: Make your contribution necessary; say no more than you must (given Q).[1]

The Q principle implements, according to Horn (1984), Grice's first submaxim of Quantity and his first two submaxims of Manner, while the I principle implements Grice's second submaxim of quantity, the maxim of relevance, and the remaining submaxims of manner. 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 two kinds of generalized conversational implicatures: *scalar* and *clausal* ones as formalized by Horn (1972) and Gazdar (1979). It is used to motivate the inferences from (1a) and (2a) to (1b) and (2b), respectively. The I principle allows us to infer, from the use of an expression, its most informative or stereotypical interpretation.[2] 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 is walking, if Mary is talking', (3a), as the biconditional 'John is walking if and only if Mary is talking', (3b).

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. Some of these problems suggest that the resulting theory is too general, while others suggest that it is not general enough. I will briefly discuss these problems in this section.

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. I will concentrate mainly on generalized *scalar* implicatures, implicatures based on a context-independent scale of alternative expressions of the same grammatical category which are ordered

[1] Horn calls the second principle the R-principle.

[2] Although it remains very unclear what exactly this underlying principle really amounts to. I have seen no proposed implementation of the I principle that accounts for all inferences that it is supposed to take care of.

by informativity, or semantic strength. The idea is that ⟨all, some⟩ is an acceptable scale, because (1c) *entails* (1a), but not vice versa.

(1a) John ate some of the cookies.

(1c) John ate all of the cookies.

From this scale, plus the fact that (1a) was used, we can then conclude that (1c) is not the case, because otherwise the speaker would have said so. In general, if the speaker asserts that a lower point on the scale obtains, he implicates that the stronger point does not obtain. In his classic work on implicatures, Gazdar (1979) proposes that in such cases, we can conclude that the speaker *knows* that the stronger point does *not* obtain. But this strong proposal, in combination with the assumption that *entailment* is a sufficient condition for being an alternative on a scale, gives rise to serious problems.[3]

First, a Gazdarian analysis of scalar implicatures predicts that from my assertion (4a), you can conclude that (4b) is not the case:

(4) a. Someone is sick.

b. John is sick.

Since (4b) entails (4a) the theory predicts that I know that (4b) is not true, and thus, that it is in fact false. But, because this could be concluded for every individual, an inconsistency arises. Gazdar gets rid of this implication by saying that in this case the potential implicature is *cancelled*: this potential implicature is inconsistent with a *clausal* implicature triggered by the disjunction used in (4a). However, this is not enough. We would also like to predict that from a disjunctive sentence of the form *John or Mary or Sue is sick* we can conclude that only one of the three is sick. Assuming that disjunction is an n-ary connective, Gazdar predicts only that not all the disjuncts are true, which is too weak.[4]

Second, because *iff* is stronger than *if*, ⟨iff, if⟩ forms a legitimate scale, and it is predicted that the falsity of (3b) is a scalar implicature of the use of (3a).

(3a) John walks, *if* Mary talks.

(3b) John walks *iff* Mary talks.

In fact, however, we typically *do* infer from (3a) that (3b) is the case. This latter intuition is accounted for by the generated I-implicature which predicts that (3b) *does* follow from (3a). But the Q- and I-implicatures involved are in direct conflict with one another, which is problematic.

[3] Atlas & Levinson (1981) suggest replacing the entailment relation by a more general informativity relation, *inf*, as defined by Carnap & Bar Hillel (1952). This generalization does not affect the discussion in this section, however.

[4] As Arthur Merin (p.c.) showed me, we can predict better if we assume that disjunction is always binary. However, I don't think we should treat 'Someone' as a complex term involving a binary connective.

In defense of neo-Gricean pragmatics, Soames, Horn, Levinson and others have proposed various solutions to these problems. It turns out, unfortunately, that there are problems with all of their suggestions.

To account for the first problem, Soames (1982) has proposed that we should *weaken* the force with which the generalized scalar implicatures are generated: from a sentence of the form (4a) we cannot in general conclude that the speaker knows that the stronger (4b) is not the case, but only that the speaker *doesn't know* whether (4b) is the case. Indeed, in this way we get rid of the problematic prediction discussed above, but now it becomes unclear how to account for the *exclusive* reading of a *disjunctive* sentence (cf. Groenendijk & Stokhof, 1984). It seems that precisely to account for the latter, Gazdar proposed to generate such strong scalar implicatures. Perhaps, then, the exclusive reading of *or* should (somehow) be accounted for in terms of a *particularized* conversational implicature,[5] but we will see that this weaker commitment of implicatures is problematic in other domains.

To account for the second problem and some others, one can put constraints on what counts as a contrastive expression. Entailment is then no longer assumed to be a *sufficient* condition for formulating scales.

One important constraint, proposed, among others, by Gazdar (1979) and Atlas & Levinson (1981), and motivated by the Gricean submaxim of Brevity, demands that the alternative expressions of a scale must be lexicalized to the same degree. By restricting the alternatives of lexical expressions to other lexical expressions one can easily account for the fact that one cannot conclude by scalar implicature that '*not(B iff A)*' follows from '*B, if A*'. However, this proposal seems to be too strong. First, as argued for by Matsumoto (1995), one way to account for the intuition that we can infer (5b) from (5a) is by scalar implicature.

(5) a. Mary caused John to die.

b. Mary killed John indirectly, e.g., by poisoning him.

For this implicature to go through, we have to assume that the lexical causative *kill* has a stronger semantical meaning than the periphrastic causative *cause to die* by being semantically restricted to stereotypical causation, and that ⟨kill, cause to die⟩ forms a legitimate scale. As noted by Matsumoto (1995), however, this latter assumption is in conflict with the above proposed constraint on appropriate scales.[6,7]

[5] Indeed, this is what Soames (1982) proposes. Following Horn (1968), Soames proposes that only in cases the speaker can be presumed to know the truth value of the stronger items of the scale, the implicature with the Gazdarian strong commitment follows.

[6] Matsumoto (1995) himself proposes to rule out a scale of the form ⟨iff, if⟩ by the *monotonicity condition* that expressions of a scale must be either all positively scalar or all negatively scalar. The reason is that *iff* is non-monotonic.

[7] It should be noted, though, that Horn (1984) doesn't account for the fact that (5a) gets reading (5b) via a scalar implicature. He doesn't even think that this is possible, because he denies that lexical causatives have a

There are also other arguments against the proposed restriction on alternatives. Suppose that we follow Soames' (1982) suggestion that scalar implicatures are *weak* in the sense that we can conclude only that the speaker doesn't know that the stronger item is true. This analysis seems to allow us to account for Grice's (1989) observation that we can normally infer (6c) from A's answer (6b) in terms of a scalar implicature:

(6) a. Where does C live?

b. A: Somewhere in the south of France.

c. A does not know which town C lives in.

However, for this inference to go through we have to assume that names of specific towns are alternative expressions of (6b). But this is not allowed according to the proposed limitation on alternative expressions.

We can conclude that it is far from easy to state an appropriate proposal to put limitations on what counts as an alternative expression. Indeed, the major problem for an appropriate formulation of implicatures in terms of the Q principle is to find additional constraints on what counts as a good alternative expression: without suitable restrictions, the principle will overgenerate enormously.

However, even if we find a successful constraint on what counts as an alternative expression, the Q-principle by itself will still generate too many implicatures. According to the standard analysis, Q-implicatures are thought of as *generalized* conversational implicatures (PCIs) triggered by specific lexical items. This means that if the two expressions W and S form a scale, ⟨S, W⟩, a (non-complex) sentence in which the expression W occurs will always trigger the implicature that the corresponding sentence where S is substituted for W is not true. In some contexts, however, this will give rise to wrong predictions. This problem is standardly discussed for *numeral* expressions. Based on the assumption that numerals get an *at least* interpretation, Horn (1968) assumes that they form scales like $\langle ..., 4, 3, 2, ...\rangle$. However, the existence of such a scale would falsely predict that A's answer (7b) to Q's question (7a) implicates that (7c) is true (cf. Kempson, 1986; Van Kuppevelt, 1996).

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

b. A: John has 2 children.

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

stronger *semantic* meaning than their periphrastic alternatives. So, our argument against the proposed restriction on alternatives of a scale doesn't seem to bite him. Still, his own explanation appeals crucially to *blocking* of a stronger interpretation of (5a) due to the interaction of the Q and I principle: the stronger interpretation is blocked by the Q principle because this is already I-implicated by the lexical causative, i.e., by an *alternative expression.* But, then, even for this explanation we have to assume that *kill* and *cause to die* are alternative expressions for one another, something that is predicted to be impossible once the Q principle allows only for alternative expressions lexicalized to the same degree.

The *at least* interpretation of numerals has been widely disputed (e.g. Carston, 1995, 1998), however. As noted already by Horn (1972), this interpretation leaves it unclear how to account for the *at best* reading of the numeral in *John can run the 100 meters in 10 seconds.* But the phenomenon is not restricted to numerals. Due to the ⟨and, or⟩ scale, it is standardly assumed that we can Q-implicate (8b) by saying (8a):

(8) a. There are cookies or chocolates in the box.

b. There are cookies in the box, or chocolates, but not both.

However, this inference is not always allowed. In particular this is not the case when (8a) is given as an affirmative answer to the following *yes/no*-question:

(9) Are there cookies or chocolates in the box?

These examples seem to suggest that 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 at least for (7b) and (8a), in such *particular* conversational situations, the *generalized* conversational implicatures (8b) and (7c) are *cancelled* for reasons of relevance: the answers are already *informative enough* for the purpose of the conversation. But why should we even *trigger* implicatures for reasons of informativity to be cancelled later for reasons of relevance? Everything else being equal, it would be strongly preferred to have a theory that can do without canceling for reasons of relevance. Following Hirschberg (1985), I will propose in this paper that 'quantity' implicatures are topic-dependent, and that – for reasons of relevance in these particular situations – the (potential) implicatures do not even arise.[8] Notice that as a consequence, it is predicted that implicatures depend on the *particular* conversational settings in which they are used.

2.2.2 Not general enough: complex sentences

Not only does the standard analysis of Quantity implicatures overgeneralize, it also doesn't seem to be general enough. First, Gazdar (1979) notes that, in many conversational situations, sentence (10) implicates that the speaker doesn't know the truth value of ϕ:

(10) Aquinas said that ϕ.

It is not clear, however, how to account for this implicature in terms of standard Gricean informal reasoning, for ϕ does not entail (10).

[8] See also van Kuppevelt (1996), Scharten (1997), Carston (1998), and Merin (1999) for analyses in the same spirit. However, our analysis will be much more formal than the first three, and only Carston intends to give an analysis as general as ours. Notice that an analysis of conversational implicatures in which there is no room for cancellation is very much in the same spirit as the popular satisfaction theory of presuppositions as developed by Stalnaker, Karttunen and Heim. This analysis of presuppositions, just as my analysis of implicatures to be presented below, is an alternative to Gazdar's (1979) cancellation-based analysis of pragmatic inferences.

Gazdar (1979) proposes a projection method for how *complex* sentences inherit the implicatures of their parts. He proposes that normally an implicature of a simple sentence also be an implicature of the complex sentence of which it is part. Unfortunately, however, Gazdar's formalization overgenerates again. Soames (1982) observes that (10) implicates that the speaker doesn't know the truth value of ϕ only in particular circumstances: when the conversational participants are trying to determine the truth value of ϕ, and not when the speaker wants to present a historically accurate account of the views of Aquinas. Indeed, as Soames suggests, in the latter kinds of contexts, determining the true value of ϕ may be *irrelevant* to the conversation.

Gazdar's (1979) projection method for implicatures seems to give rise to other problems, as well. First, it is unclear what to do in case a sentence contains more than one scalar expression. The following straightforward analysis might be suggested: substitute all weak scalar items occurring in a sentence for a stronger alternative of the scale and implicate that the resulting sentence is false. This would have the effect that the two occurrences of 'some' in (11a) are replaced by 'all', resulting in implicature (11b).

(11) a. Mary ate some of the bacon and some of the eggs.

b. It is not the case that Mary ate all of the bacon and all of the eggs.

c. Mary didn't eat all of the bacon and she didn't eat all of the eggs.

Chierchia (ms), however, rightly argues against the above suggested analysis: the implicature (11b) is too weak. Sentence (11a) rather implicates (11c). There are other reasons why complex sentences are problematic. We have to explain why (12a) doesn't give rise to the implicature that (12b) is false:

(12) a. If John stole a pear, he will be punished.

b. John stole all pears.

Gazdar (1979) proposes to account for this fact by demanding that a complex sentence generates the scalar implicatures of its parts due to scale $\langle S, W\rangle$ only if the complex sentence as a whole entails the sentential clause in which the weaker expression occurs. Because (12a) does not entail that John stole a pear, the implicature that (12b) is false is correctly predicted not to arise. However, as observed by Atlas & Levinson (1981), this doesn't describe all our intuitions for downward entailing contexts. We also have to account for the fact that, in negative contexts, the scalar implicatures seem to be *reversed.* It seems that we can conclude from (13a) and (14a) that (13b) and (14b) are false, respectively:

(13) a. John didn't eat all of the pears.

b. John ate none of the pears.

(14) a. It's not the case that Rick is both a philosopher and a poet.

b. Rick is neither a philosopher nor a poet.

Atlas & Levinson propose to generalize the Horn/Gazdar analysis by suggesting that negation *reverses* scales: if ⟨all, some⟩ and ⟨and, or⟩ form Horn scales, so do ⟨none, not all⟩ and ⟨neither nor, not both⟩,[9] and in negative contexts the 'at least' interpretation of numerals should be replaced by their corresponding 'at most' interpretation. However, we also have to account for the intuition that (15a) and (16a) give rise to the inferences (15b) and (16b) that go in opposite directions:[10]

(15) a. I can run the 100 meters in 9 seconds.

b. I can't run the 100 meters in 8 seconds.

(16) a. I can jump 8 meters high.

b. I can't jump 9 meters high.

These examples strongly suggest that 'scalar' implicatures should be calculated at the *global* level: it seems that the reason why the numbers go *up* in the former case and *down* in the latter is that, in the former case, the whole sentence (15a) forms a context C such that $C(8)$ *entails* $C(9)$, while in the latter the sentence forms a context C' for which it holds that $C'(9)$ *entails* $C'(8)$.[11]

So it seems that to account for scalar implicatures, entailment should play an important role. The previous examples, however, suggest that entailment should not be used at the *local* level to determine a *context independent* scale of lexical items, but rather at the *global* level and take the whole sentence into account.[12]

Although entailment seems relevant for the analysis of scalar implicatures, we have seen in the previous section that entailment is not a *sufficient* condition for defining scales. As observed by Fauconnier (1975), Gazdar (1979) and Soames (1982), but stressed by Hirschberg (1985), entailment is not a *necessary* condition either. It seems that neither entailment nor a more general notion of 'being more informative' (e.g. the one of Carnap & Bar Hillel, 1952) is needed/wanted to account for certain examples that according to Hirschberg should be analyzed as scalar implicatures. One possible context-dependent scale not reducible to entailment is one of autographic prestige, which is involved in the inference from (17b) to the falsehood of (17c).

(17) a. Q: Whose autograph did you get?

[9] However, as Atlas & Levinson (1981) observe, the reverse implicature does not always follow: *It's not the case that Kurt went to the store and bought some wine* does not implicate that *Kurt went to the store or bought some wine.* The reason is, or so they argue, that in this case the two 'belong together': the former is done *in order* to do the latter.

[10] Although *both* (15a) and (16a) are somewhat exaggerated.

[11] See Zeevat (1994).

[12] Compare this with Fauconnier's (1975) and Krifka's (1995) analyses of NPI licencing. Both seek to reduce this phenomenon to scalar implicature, and in both accounts the informativity of *the whole sentence* in which the negative polarity item occurs plays a crucial role.

b. A: I got Joanne Woodward's.

c. A: I got Paul Newman's.

The inference is based, let us say, on the following scale ⟨Newman, Woodward, actress X⟩ with exciting Paul Newman on top, and dull actress X at bottom. The intuition is that an answer like (17b) makes clear that higher items are not true, but the answer doesn't seem to rule out that lower items are true as well. The problem for the standard analysis is that although the inferences involved are clearly 'scalar' in nature, it is hard to imagine how such a scale could be defined in terms of a notion of informativity.[13]

Atlas & Levinson (1981) observe that an assertion like (18a) is typically interpreted as meaning (18b):

(18) a. Russell wrote "Principia Mathematica".

b. Russell wrote "Principia Mathematica" by himself.

Although this inference does not directly seem to depend on an ordered scale, still Hirschberg (1985) argues that it should be accounted for in terms of Grice's maxim of quantity. Atlas & Levinson claim something similar, but they propose that the implicature should be accounted for in terms of the I principle rather than the Q principle. Carston (1998) rightly wonders: Why should we suddenly use the I principle in this case, instead of the Q principle? Atlas & Levinson argue that this is so because the Q-implicature is *cancelled* by being incompatible with the world-knowledge that books are typically written by a single author.

Perhaps this (rather ad hoc) analysis can account for the case above. It can't be the whole story, though, because the phenomenon to be explained is much more general. As Groenendijk & Stokhof (1984) observe, when A answers Q's question (19a) by saying (19b), we typically interpret the answer as being exhaustive.

(19) a. Q: Who came to the party?

b. A: John came.

That is, we interpret A's answer exhaustively as "*Only* John came": the assertion of one alternative implicates the denial of the others. They also note that, although on first thought, this kind of inference should intuitively be accounted for in terms of Grice's maxim of Quantity (as a Q-implicature), the standard implementation rather predicts the opposite. From the existence of the informativity scale ⟨Only John, John⟩ we should conclude, via the Q principle, from the answer (19b) that other people besides John came to the party as well.[14] As we have seen above, however, it is not clear on what grounds this opposite Q-implicature could be cancelled.

[13] For a discussion of some other problematic examples, see Hirschberg (1985) and Scharten (1997).

[14] See Atlas & Levinson (1981) for similar reasoning.

3 Implicatures and Exhaustivity

In the previous section I have suggested several times that implicatures very much depend on the issue involved. We approve of Soames's (1982) suggestion, for instance, that from (10) we can conclude that the speaker doesn't know the truth value of ϕ only if the truth of ϕ is at issue. In this section I will explain how most of the 'quantity' implicatures discussed above can be explained if we assume that assertions are exhaustified answers to questions.

3.1 Exhaustify term-answers

Groenendijk & Stokhof (1984) propose to account for the intuition that answer (19b) to question (19a) should normally be read exhaustively by introducing an explicit exhaustivity operator that is applied to the denotation of the term given in the answer to derive the exhaustive interpretation. In this section I will follow them and show that almost all typical quantity implicatures can be alternatively analyzed on the assumption that assertions are exhaustified answers to questions. We will see that this alternative analysis can account for certain inferences that the standard one based on the Q and I principles cannot. Moreover, the analysis does not give rise to the above discussed overgeneralizations, because the implicatures are made *topic-dependent*.

Consider the terms (21a)-(21c) given as answers to question (20):

(20) Who is sick?

(21) a. John.

b. John and Bill.

c. A man.

Groenendijk & Stokhof (1984) want the answers to mean *Only John*, *Only John and Bill* and *Only a man*, respectively. They propose that this can be done by applying the following exhaustivity operator to the terms denoted by (21a)-(21c):

$$\begin{array}{lcl} \underline{exh}^{GS} & = & \lambda T\lambda t\lambda w[T(t)(w) \wedge \neg\exists t'[T(t')(w) \wedge t'(w) \neq t(w) \wedge \forall x[t'(x)(w) \rightarrow t(x)(w)]]] \\ \underline{exh}^{GS}(T)(t) & = & \{w \in T(t) : \neg\exists t'[T(t')(w) \wedge t'(w) \subset t(w)]\} \end{array}$$

In this formula, t stands for the property underlying the *wh*-question, for (20), for example, for a function from worlds to the set of individuals that are sick in that world. T stands for the denotation of the term-answer. In (21) for a property of properties, i.e. in each world a set of sets of individuals. Set-theoretically, the above formula just picks out the *minimal* elements of a set of sets. Let us assume that our domain contains 3 individuals: John, Bill, and Mary. In that case, the term John, $\lambda P.P(j)$, corresponds to the following generalized quantifier: $\{\{j\},\{j,b\},\{j,m\},\{j,b,m\}\}$. The formula resulting from applying $\underline{exh}^{GS}$ to $\lambda P.P(j)$ is $\lambda P.\forall x[P(x) \leftrightarrow x = j]$, with $\{\{j\}\}$ as the corresponding quantifier. If (21a) is then given as the true exhaustive answer to the question, we can conclude that $\{j\}$ is the set of individuals that are sick.

Similarly, we can see that the resulting formulae and sets for (21b) and (21c) are $\lambda P \forall x[P(x) \leftrightarrow [x = j \vee x = b]]$ corresponding with $\{\{j, b\}\}$ and $\lambda P \forall x[P(x) \leftrightarrow x = j] \vee [P(x) \leftrightarrow x = b]]$ corresponding to $\{\{j\}, \{b\}\}$, respectively. Thus, if we assume that (21c) is given as the true exhaustive answer to question (20), we can conclude that either only John is sick or only Bill.

3.2 Scalar implicatures for terms

This exhaustivity operator accounts for many of the implicatures traditionally accounted for in terms of Grice's maxim of quantity.[15] First, it obviously accounts for the fact that when (20) is answered by (21a) we conclude that *only* John is sick. Second, when answer (21c) is given we can conclude that not all men are sick, an inference standardly triggered by the ⟨all, some⟩ scale. Recall that when we ignore clausal implicatures, Gazdar's account would predict that John is not sick. We don't make this incorrect prediction. The exhaustive reading of (21c) as answer to question (20) is compatible with the fact that John is sick.

We have seen in section 3 that there has been a lot of discussion about the strength with which scalars should be implicated. Gazdar (1979) proposed that when a weak item of a scale is used, we can conclude that the speaker *knows* that the stronger items are *not* true. Soames (1982), on the other hand, proposed that in such situations we can conclude only that the speaker *doesn't know* that the stronger items are true. Both positions are problematic, however. Gazdar generates implicatures that are *too strong*, for it wrongly predicts that from answer (21c) to question (20) we can conclude that John is not sick. Soames's proposal seems to be *too weak*, for it remains unclear how to account for the fact that disjunctions are typically read exclusively, and why from answers (21a) and (21c) above we can conclude that it is not the case that both John and Mary are sick.

When we assume that 'implicatures' are generated via exhaustification, however, we immediately specify the epistemic force of the implicatures in a more appropriate way: from a speaker's assertion/answer we can implicate that the speaker knows that its exhaustified reading is true, but that she doesn't know anything more. Thus, the answer (21a) to question (20) implicates that the speaker knows that no one else but John is sick, while answer (21c) implicates that the speaker knows that no woman is sick, but that one and only one man is, though she doesn't know which one. We will see soon that this assumption of epistemic commitments associated with assertions also predicts that in lots of cases where a speaker uses a disjunctive sentence, she knows that one and only one of the disjuncts is true, but does not know which one. So, we don't need Gazdar's (1979) rather ad hoc formulation of scalar and clausal implicatures to account for these intuitions. Moreover, our analysis immediately extends to n-ary disjunctive terms, while Gazdar's does not.

[15] This is not only the case for Groenendijk & Stokhof's exhaustivity operator, but also for Zeevat's (1994) appealing and very Gricean alternative. For an extensive discussion of the workings of several exhaustivity operators – including the one used in section 4 of this paper – see van Rooy & Schulz (ms).

Chierchia's (ms) problem for the traditional analysis of scalar implicatures also does not arise. 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 eat?*, we can conclude that Mary didn't eat all of the bacon, and that she didn't eat all of the eggs, just as Chierchia claims we should. The reason is that the term should be represented as something like $\lambda P[\exists x[B(x) \wedge P(x)] \wedge \exists y[E(y) \wedge P(y)]]$.[16] Assuming for simplicity that there are (only) two pieces of bacon and eggs each, we predict that, after exhaustification, the term denotes a set like $\{\{b_1, e_1\}, \{b_1, e_2\}, \{b_2, e_1\}, \{b_2, e_2\}\}$. As a result, the answer is predicted to mean that Mary ate no more than some (one piece) of the bacon and some (one piece) of the eggs.

In terms of exhaustification we can account for implicatures arising in other kinds of complex sentences as well. As noted by Chierchia (ms), we typically infer (22b) from a sentence like (22a).

(22) a. John believes that some men are sick.

b. John believes that not all men are sick.

It should be obvious that this inference is accounted for straightforwardly when (22a) is given as an answer to (perhaps implicit) (23a). However, I believe that the inference will also typically arise in case (22a) is be used to answer question (23b) (probably with contrastive accent on 'John').

(23) a. Who does John believe is sick?

b. Who is sick?

In the latter case, the answer can be relevant if, for instance, John is thought to be a specialist on the issue of who is sick.[17]

As it turns out, our exhaustification analysis accounts not only for cases standardly accounted for in terms of the Q principle, but also some inferences standardly justified by the I principle. Exhaustification correctly predicts, for instance, that (18a) is typically interpreted exhaustively as meaning (18b) when given as an answer to the (perhaps implicit) question *Who wrote Principia Mathematica?*, and no conflict between Q and I implicatures is generated. More interestingly, perhaps, is that exhaustivity also helps us account for the phenomenon of *conversion*, the fact that we sometimes infer (24b) from (24a):

(24) a. Every duck quacks.

b. Every quacker is a duck.

[16] I am now, as in the rest of the paper, ignoring the singular-plural distinction. It is important to note, though, that Groenendijk & Stokhof's exhaustification operator works equally straightforward to plural noun phrases.

[17] My suggestion that pragmatic inferences of *embedding* sentences can be due to an issue arising at the *main* level bears a close resemblance to Heim's (1992) observation that the presupposition trigger *too* in the embedded clause of (ii) can be used appropriately when (i) has been stated just before:

(i) Sue will come to the party.

(ii) John believes that Mary will come *too*.

Recently, this phenemenon has been extensively discussed in Horn (2000), where he suggests accounting for it in terms of an *I*-implicature. How the inference should be accounted for exactly, however, Horn did not inform us. It is easy to see how the inference can be accounted for in terms of exhaustivity. The reason is that from the exhaustive interpretation of term-answer *Every duck* to the question *Who quacks?*, we can conclude that (24b) is the case.

Our approach also predicts that (25a) should implicate (25b) when the color of the flag is at issue.

(25) 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 (25a) should be interpreted as being as informative as possible, i.e., as inducing an *I*-imlicature. But, it should then 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 (red, yellow, or green). Since the color of the patch distinguishes the flag, if you ask me to identify which of the flags you hold behind your back, my answer (25a) satisfies you, and does not imply (25b). The standard analysis has to assume that, in these cases, the triggered generalized implicature is *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 flag's patch?*

Indeed, our topic-dependent analysis of 'scalar' implicatures prevents us from triggering implicatures to be *cancelled* later for reasons of relevance. Consider the following example again:

(7) 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 (7b) triggers a potential implicature (7c) that is cancelled when the former is given as an answer to question (7a), our analysis predicts that the implicature is not even triggered, *because* (7b) completely answers (7a). We will see later that a similar analysis can be given for the fact that a disjunctive sentence sometimes indeed gets a non-exclusive reading.[18]

One might object to our approach, saying that scalar implicatures arise even when a scalar term is not used to answer a corresponding question. I believe this objection is ungrounded.

[18] This suggests that our exhaustivity analysis predicts that implicatures will never be cancelled. However, we also have to deal with answers like *John or Bill or both* and *At least John* to question *Who are sick?* On the assumption that these answers have the same semantic meaning as the alternatives without the extra *or both* and *at least*, an exhaustivity analysis has to allow for cancellation. However, Groenendijk & Stokhof (1984) suggest that when we take *plurality* into account the answers can be predicted to have a different meaning, and Bonomi & Casalegno (1993) show how this can account for the non-cancellation effect. Though appealing, I am not fully convinced of this proposal; perhaps cancellation should be allowed for when extra *effort* is used.

Assuming the standard scale ⟨know, believe⟩, this would give rise to the incorrect prediction that (26b) implicates (26c) if used to answer (26a).

(26) a. Who believes that Columbus discovered America?

b. John (believes that Columbus discovered America).

c. John does not know that Columbus discovered America.

Thus, our exhaustification approach towards implicatures predicts that they depend on the topic being addressed. This topic can be an explicitly stated question or an implicit issue that is somehow relevant in the discourse (as in the case of Grice's (1967) example of the man with the obviously immobilized car). An implicit issue can be made explicit by the use of accent in the assertion. Lot's of work has been done recently about how the use of accent indicates which (implicit) topic is being addressed (e.g. Rooth, 1992; Roberts, 1996), but I won't go into that here.

3.3 Generalized exhaustivity

Until now we have used the simple exhaustivity operator for quantified terms in extensional contexts, and shown how it can account for the so-called 'scalar' implicatures. Groenendijk & Stokhof (1984) show that the above stated operator for terms can be generalized easily to multiple terms and sentential answers. I will leave multiple terms for what they are and consider only 0-ary interrogatives, i.e. *polar questions*. In that case the answers are considered to be *sentential modifiers*, functions from propositions to propositions. The answer *yes*, for instance, will be interpreted as $\lambda p \lambda w[p(w)]$, and *If Mary talked* as $\lambda p \lambda w[T(m)(w) \rightarrow p(w)]$.

$$\begin{array}{lcl} \underline{exh_0}^{GS} & = & \lambda T \lambda p \lambda w[T(p)(w) \wedge \neg \exists p'[T(p')(w) \wedge [p'(w) \neq p(w)] \wedge [p'(w) \rightarrow p(w)]]] \\ \underline{exh_0}^{GS}(T)(p) & = & \{w \in T(p) : \neg \exists p'[T(p')(w) \;\&\; p'(w) \neq p(w) \;\&\; (\neg p'(w) \vee p(w))]\} \end{array}$$

Applying this operator to answers which contain *if*'s and *or*'s, Groenendijk & Stokhof notice that (27c), meaning (27d), results as the exhaustive interpretation of answer (27b) to question (27a), and that similarly (28b) is read as (28d) when interpreted as the exhaustive answer to question (28a):

(27) a. Q: Did John walk?

b. A: *If* Mary talked.

c. $[T(m)(w) \rightarrow W(j)(w)] \wedge \neg \exists p[[T(m)(w) \rightarrow p(w)] \wedge [p(w) \neq W(j)(w)] \wedge [p(w) \rightarrow W(j)(w)]]$

d. John walked *iff* Mary talked.

(28) a. Q: Are the cookies in the box?

b. A: Yes, *or* the chocolates.

c. $[IB(c)(w) \vee IB(ch)(w)] \wedge \neg\exists p[[p(w) \vee IB(ch)(w)] \wedge [p(w) \neq IB(c)(w)] \wedge [p(w) \rightarrow IB(c)(w)]]$

d. Either the cookies or the chocolates are in the box, but not both.

However, this analysis does not have the result that *if* should always be read biconditionally, nor that *or* should always have the exclusive reading. In particular this is rightly predicted not to be the case in (29) and (30), where the complex sentences are given as (exhaustive) polar answers:

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

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

(30) Q: Are there any cookies or chocolates in the box?

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

3.4 Problems for Groenendijk & Stokhof's exhaustification

Unfortunately, as noted by both Groenendijk & Stokhof (1984) themselves, their exhaustivity operator does not always give the right results. This is in particular the case for downward monotone quantifiers like *Not John* and *No girls*. Applying the exhaustivity operator to answer (31b) results in the wrong prediction that the answer really means the same as (31c):

(31) a. Q: Who is coming to the party?

b. A: No girls.

c. A: No one.

The reason is that the proposed exhaustivity operator picks out the *minimal* elements of a quantifier, and for monotone decreasing quantifiers this is always the empty set. Groenendijk & Stokhof (1984) propose to account for this problem by assuming that a phrase like "At most two girls" gets a 'plural' interpretation and denotes a set like

$$\{X | \ \{G\} \subseteq X, \text{ where } G \text{ is a group of girls having at most 2 members}\}$$

The smallest elements of this set, picked out after exhaustification, are indeed sets of girls with at most 2 members. This analysis is somewhat counterintuitive, however, for three reasons: First, it forces a collective interpretation of terms, even though the predicates might be distributive. Second, it cannot account for the phenomenon of *scale reversal*: that from the assertion of *John didn't eat all of the apples* we normally can conclude that John ate at least some of the apples, as long as we allow the empty group to be the smallest set of individuals. Third, even if "At most two girls" is explicitly marked as a partial answer and thus that exhaustivity does not apply, the answer should be false if, for example, three girls are coming to the party. This, however, is not predicted to be the case in the above suggested analysis. For these reasons we won't adopt their solution.

A second problem with Groenendijk & Stokhof's exhaustivity operator is that it predicts incorrectly for so-called mention-some questions. Sometimes an assertion intuitively answers a question completely satisfactorily without being read exhaustively in Groenendijk & Stokhof's sense. To illustrate, when I ask you (32a) and you answer with (32b), 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.

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

b. Around the corner, at Oude Gracht 15.

Thus, the notion of exhaustivity should take into account that an answer can intuitively *resolve* a question without ruling out other possible resolving answers.

Third, Groenendijk & Stokhof's (1984) analysis cannot account for implicatures involving, for instance, the autographic prestige scale as discussed in section 3. Although the claim that I got an autograph of Joanne Woodward intuitively rules out that I got an autograph of exiting Paul Newman, it doesn't seem to rule out the fact that I got an autograph of dull actress X. This intuition cannot be accounted for simply by assuming that the claim is the G&S-exhaustive answer to the question *Which movie stars' autographs did you got?*.[19] Thus, this type of example suggests that Groenendijk & Stokhof's exhaustivity can't do all of the work, and that we still need scales. These scales, however, cannot in general be defined in terms of informativity.[20]

Finally, even for the 'quantity' implicatures that can be accounted for by assuming that these sentences should be interpreted as exhaustive answers to questions in Groenendijk & Stokhof's sense, we would like to say something more. We would like to find a more general reason for *why* answers should normally be interpreted exhaustively, an answer which relates the analysis closer to scalar analyses. Thus, contrary to Groenendijk & Stokhof (1984), we would like to *derive* the fact that only the minimal elements of the quantifiers expressed by term-answers are selected, and in such a way that if we look at things more generally we can (i) explain why some questions and answers get a mention-some interpretation; and (ii) give a satisfactory account of, for instance, scales of autographic prestige. In the next section I will show how we can do this in terms of a principle called *maximization of relevance.*

4 Relevance and Implicatures

It is widely acknowledged in the literature that standard *particularized* conversational implicatures are topic-dependent; the implicatures are derived from the fact that it is common knowledge that

[19] Unless the question itself should have a 'scalar' reading. This latter possibility is allowed for in my relevance-based analysis of questions (van Rooy, ms).

[20] Bonomi & Casalegno (1993) argue similarly that sentences in which 'only' occurs don't just have a standard exhaustive reading, but might have a separate scalar reading, too.

a certain utterance is made to help to resolve a question, or decision problem, of one or more of the participants in the discourse. These examples are normally explained by appealing to Grice's informal principle that says: *Be relevant.* In the following subsections I will show that most of the implicatures discussed in this paper can be accounted for by the somewhat stronger principle which says that one should be *maximally relevant* (cf. Sperber & Wilson). But in order to determine which interpretation is maximally relevant, we need to be able to order these interpretations by relevance. In the following subsection I will propose a way of doing so.

4.1 Topic-dependent relevance

According to Groenendijk & Stokhof (1984), a question like 'Which individuals have property P?, represented as '$?xPx$', gives rise to an equivalence relation and should be analyzed by means of the following lambda term or its corresponding partition:

$$\begin{aligned} [[?xPx]]^{GS} &= \lambda w \lambda v[\lambda x P(w)(x) = \lambda x P(v)(x)] \\ &= \{\{v \in W | \ \forall d \in D: \ d \in P(v) \text{ iff } d \in P(w)\} | \ w \in W\} \end{aligned}$$

The idea is that the meaning of a question is its set of resolving answers, and that to give a true resolving answer in a world you have to give the *exhaustive* list of individuals that have the relevant property in this world. Thus, for John to know the answer to the question *Who is sick?*, for instance, John must know of *each* (relevant) individual *whether* he or she is sick. To give a true and resolving answer to the question, he must *mention all* the individuals that are sick, and implicate that this is indeed the whole list. This partition analysis predicts that two worlds are in the same cell of the partition induced by the above question, if the property *being sick* has the same extension in both worlds. The elements of the partition $Q = \{q_1, ..., q_n\}$ are called the *complete answers.* An assertion counts as a *partial answer* to a question iff the proposition it expresses is incompatible with at least some but not all cells of the partition. Thus, complete answers are special kinds of partial answers; they are the most informative kinds of partial answers and are true in the worlds of just one cell of a partition. This suggests, perhaps, the following proposal: say that one answer is 'better' than another, just in case the former *entails* the latter. But this would be mistaken, for it would wrongly predict that we prefer *over*informative answers to answers that are just complete. If I ask you, for instance, whether John is sick, I would be very puzzled by your answer *Yes, John is sick, and it is warm in Africa.* The second conjunct to the answer seems to be completely *irrelevant* to the issue, and thus should not be mentioned. So it seems that we should measure how good an answer is mostly in terms of the partition induced by the question. And, indeed, this is exactly what Groenendijk & Stokhof (1984) propose.

Define A_Q as the set of cells of partition Q that are compatible with answer A:

$$A_Q \quad = \quad \{q \in Q: \ q \cap A \neq \emptyset\}$$

Notice now that one partial answer A can be more informative than another one B because it is incompatible with more cells of the partition than the other one, i.e. $A_Q \subset B_Q$ (where '$\subset$' is proper set inclusion). Groenendijk & Stokhof propose that when answer A is incompatible with more cells of the partition than answer B, i.e. $A_Q \subset B_Q$, the former should be counted as a better answer to the question than the latter.

But what if two answers are incompatible with the same cells of the partition, i.e. if $A_Q = B_Q$? It is possible that when two partial answers to a question that are incompatible with, for example, just one cell of the partition, one of them can be more informative than the other because the former *entails* the latter. As we have already suggested above, the latter counts in that case as a better answer than the former, because it doesn't give extra *irrelevant* information. Thus, in case $A_Q = B_Q$, A is a better answer than B iff $A \supset B$. Combining both constraints, Groenendijk & Stokhof (1984) propose that A is (quantitatively) a *better* semantic answer to question Q than B, $A >_Q B$, by defining the latter notion as follows:

$$\begin{array}{llll} A >_Q B & \text{iff} \quad \text{either} & \text{(i)} & A_Q \subset B_Q, \text{ or} \\ & & \text{(ii)} & A_Q = B_Q \text{ and } A \supset B. \end{array}$$

I will interpret the ordering relation '$>_Q$' as one of *relevance* and say that $A >_Q B$ iff A is a *more relevant* answer to question Q than B is. Notice that in contrast to an ordering relation based only on entailment, our ordering relation is very *context dependent.* The ordering, or *scale*, depends crucially on the background question. Notice, however, that entailment is a special case of our ordering '$>_Q$': if the question is what the *world* is like, the ordering obviously comes down to entailment.

Although Groenendijk & Stokhof's partition analysis of questions is appealing, it is not completely satisfactory. This is because it predicts that each question has at most one true resolving answer in a world. However, a question like (33) can intuitively be answered satisfactorily by mentioning just one individual, i.e. you don't have to give an exhaustive list of persons that have a light.

(33) Who has got a light?

This suggests that such questions should be analyzed (roughly) as Hamblin (1973) proposed. Restricting ourselves for simplicity to single *wh*-questions and ignoring free variables, Hamblin's analysis comes down to the following:

$$[[?xPx]]^H \quad = \quad \{\lambda w[d \in P(w)] : d \in D\}$$

Notice that in case there are worlds in which more than one individual has property P, this question denotation does not give rise to a partition.[21]

Although $[[?xPx]]^H$ does not (necessarily) give rise to a partition, we can still use Groenendijk & Stokhof's ordering relation to determine how relevant an answer is. If $[[?xPx]]^H$ denotes

[21] In van Rooy (ms) I give to questions a single underspecified semantic meaning such that the actual interpretation depends only on a contextually given relevance-relation (defined in terms of decision-problems). It is shown that both Groenendijk & Stokhof's question-semantics as well as Hamblin's are special cases.

$\{\{u, w\}, \{v, w\}\}$, for instance, it will be the case that a sentence that expresses proposition $\{u\}$ will be equally as relevant as one that expresses $\{u, w\}$, but both are more relevant than one that denotes $\{u, v\}$ or $\{u, v, w\}$.

Groenendijk & Stokhof (1984) defined an interesting ordering relation between answers but didn't really make much use of it. In the next section we will see that that was a missed opportunity; by making use of this ordering relation they could have improved significantly on their own exhaustivity operator.

4.2 Relevance, exhaustivity and scalar implicatures

Remember that in the simplest case Groenendijk & Stokhof's exhaustivity operator selects the *minimal* elements of the generalized quantifier that corresponds to the NP denotation. This operator is *context-independent* in the sense that it doesn't need to look at the (question) predicate to determine those minimal elements. Now we will define a similar exhaustivity operator, but by making use of the ordering relation '$>_Q$'. Because this ordering relation is defined in terms of *sentence*-denotations, i.e. *propositions*, unlike Groenendijk & Stokhof, we have to take the predicate into account as well. In this way, we make the exhaustivity operator more context-dependent. Our exhaustivity operator will be a function whose value is a proposition and which takes as arguments a term-denotation, the (question) predicate, and the underlying question/decision problem.[22] What intuitively goes on is very similar to the procedure used by Groenendijk & Stokhof: we will select the minimal elements of the term-denotation. However, what the minimal elements are depends now on the proposition expressed and the ordering relation induced by the question. We define the new exhaustivity operator $\underline{exh}^R$ as follows:

$$\underline{exh}^R(T)(P) \ = \ \{w \in W | P(w) \in T(w) \ \& \ \neg \exists t \in T(w) : \lambda v[P(w) \subseteq P(v)] >_Q \lambda v[t \subseteq P(v)]\}$$

To illustrate the workings of this operator, look at a quantifier denoted by the noun phrase *John.* Suppose that $D = \{j, m, s\}$. In that case $[[John]] = \{\{j\}, \{j, m\}, \{j, s\}, \{j, m, s\}\}$. When the underlying question/decision problem Q is, for instance, the mention-all question *Who is sick?*, the ordering relation $>_Q$ basically reduces to the proper subset relation '$\subset$', i.e. to (one-sided) entailment. But then there is no element $t \in [[John]]$ different from $\{j\}$ such that $\lambda v[\{j\} \subseteq sick(v)] \subset \lambda v[t \subseteq sick(v)]$. Recall that proposition $\lambda v[\{j\} \subseteq sick(v)]$ means that *at least* John is sick, and thus that our exhaustivity operator now selects the element of T that give rise to the *weakest* proposition(s) compatible with the answer. As a result, $\underline{exh}^R([[John]])(sick)$ expresses the proposition that John and nobody else is sick.

Something similar happens with term-answers that express other monotone increasing quantifiers: on the assumption that they answer the mention-all question associated with the predicate,

[22] Notice that I assume that the question predicate does not fully determine the underlying question: the question could be both of the mention-all and the mention-some kind, and which one this is depends on the underlying decision problem (see van Rooy, ms).

our above rule makes similar predictions to Groenendijk & Stokhof's. However, it works differently for answers to *mention-some* questions.

Consider the term-answer *John* to a *mention-some* question like *Who has got a light?*. The proposition that expresses that Mick and Sue have a light is no better now than the one that expresses that (at least) Mick has a light. In this case, $\underline{exh}^R([[John]])(have\ light)$ expresses the proposition that *at least* John has a light. Thus, the answer *John* is predicted not to rule out that Mick or Sue might have a light. And this is in accordance with intuition. It is easy to see, again, that an answer like *A man* now does not rule out that more than one man has a light, nor that there are women who have a light. Thus, in contrast with Groenendijk & Stokhof's exhaustivity operator, ours doesn't have the same meaning as *only*, and this allows us to also apply the operator in case the term is used as an answer to a mention-some question.

A closely related difference with Groenendijk & Stokhof's account is that we predict that the exhaustive interpretation of the term *Some men* used as an answer to the question *Who will come?* doesn't necessarily mean that only 2 men will come. Even if the question gives rise to a partition, using the question-semantics of van Rooy (ms), we still predict that the fine-grainedness of this partition crucially depends on context (an underlying decision problem). In case a proposition stating that a group of at least two men will come is equally as relevant as the proposition claiming that a group of at least three (four, etc.) men will come, but less relevant than the proposition that all men will come, we correctly predict that the answer gives rise to the inference that not all men will come, but not to the much stronger prediction that no more than 2 men will come. Notice, as an aside, that no purely Gricean analysis can account for this if we assume that informativity is modeled in terms of the context-independent notion of entailment.

Just like in Groenendijk & Stokhof (1984), our exhaustification operator also works in general for n-ary quantifiers, and in particular for sentential modifiers. This is crucial for the analysis of sentential answers involving *if* or *or*. Our operator works a little bit differently in these cases than the corresponding one of Groenendijk & Stokhof. Still, the results are the same: the sentential connective *or* will in appropriate circumstances get the exclusive reading, and applying the exhaustivity operator to the answer *If Mary talked* to the question *Did John walk?* results in the bi-conditional reading *Mary talked if and only if John walked.* The reason is that of the two truth values, '0' will be selected as the minimal one, because the truth values '0' and '1' are represented as $\emptyset$ and $\{\emptyset\}$, respectivily, and for any non-trivial proposition p, $\lambda v[\{\emptyset\} \subseteq p(v)] = \lambda v[p(v)]$ decides $?p$, while $\lambda v[\emptyset \subseteq p(v)] = \lambda v[p(v) \vee \neg p(v)]$ does not.

Unfortunately, the above analysis still goes wrong for monotone decreasing quantifiers like *Not all men.* For these cases, however, we will adopt an idea of von Stechow & Zimmermann (1985). They propose making a distinction between so-called 'positive' and 'negative' quantifiers (roughly corresponding to monotone increasing and decreasing ones, respectively), and suggest a separate exhaustivity operator for negative quantifiers which does not select the minimal elements, but instead the *maximal* ones.[23] We can adopt von Stechow & Zimmermann's solution to Groenendijk & Stokhof's problem with monotone decreasing NP-denotations by exchanging the relevant order as well:

$$\underline{exh}'^{R}(T)(P) \;=\; \{w \in W | P(w) \in T(w) \;\&\; \neg\exists t \in T(w) : \lambda v[P(w) \subseteq P(v)] <_Q \lambda v[t \subseteq P(v)]\}$$

Consider now, for example, the term-answer *Not all men* to the question *Who is sick?* which – let us assume – has a mention-all reading. On the assumption that $[[Man]] = \{j, m, f\}$, we see that $[[Not\ all\ men]] = \{\emptyset, \{j\}, \{m\}, \{f\}, \{j, m\}, \{j, f\}, \{m, f\}\}$ and that $\underline{exh}'^{R}([[Not\ all\ men]])(sick)$ expresses that either John and Mick, or John and Fred, or Mick and Fred, but not all three of them are sick, which seems to be right.[24] Notice that when the underlying question has a mention-some reading, the term-answer will include all non-empty sets of men, not just the maximal ones. Again, I think this is in accordance with intuition. Similar pleasing predictions result now for other monotone decreasing quantifiers.

When we look at negative sentences, we see that our exhaustivity operator does not predict that *or* gives rise to the exclusive reading. This observation is also relevant for implicatures arising in attitude attributions. In section 3.2 we discussed how implicatures can arise from terms used in embedded clauses of belief attributions. The same analysis accounts for the exclusive reading of *or* if used in such an embedded clause. However, the disjunctive connective does not give rise to an exclusive reading for all attitudes: (34b) cannot be inferred from (34a):

(34) a. John doubts that Mary or Sue will come.

b. John doubts that only Mary will come or that only Sue will come.

This, however, can be accounted for straightforwardly when we analyze *doubt* in terms of *believe* and *negation*. It is then predicted that *doubt* gives rise to a kind of 'scale reversal' effect, and thus that we infer from *not both* to *either or*, rather than from *or* to *not and.*

In our exhaustivity operator we make crucial use of a context-dependent ordering relation between propositions, i.e. a scale. We have seen that the ordering between propositions induced by entailment is a special case of a relevance-induced scale. In the previous section we saw that if we want to account for scalar implicatures in terms of an exhaustivity operator, we have to face the problem that some scales cannot be reduced to entailment. As noted by Hirschberg (1985), a scalar implicature can be based, for instance, on a notion of autographic prestige: in a game-like situation where the one who wins is the one who has the most prestigious autograph, the proposition saying that I have an autograph of (at least) Paul Newman is *more important/better* than the one saying

[23] They propose that for monotone decreasing quantifiers we should not use $\underline{exh}^{GS}$, but rather the following one:

$$exh^{SZ} \quad = \quad \lambda T\lambda t\lambda w[T(t)(w) \wedge \neg\exists t'[T(t')(w) \wedge t(w) \subset t'(w)]]$$

Applied to a quantifier like *not all men* result in a set consisting of sets of all but one man.

[24] This accounts for the inference from (15a) to (15b) as well, if we assume that (i) numerals get an *at least* reading, (ii) the preposition *in* in 'in 9 seconds' changes an increasing quantifier into a decreasing one (from $\{A \subseteq N : \forall x \in A : n \leq x\}$ into $\{A \subseteq N : \forall x \in A : n \geq x\}$), and (iii) the predicate is distributive. I owe this idea to Katrin Schulz.

that I have an autograph of (at least) Joanne Woodward. But such examples can be accounted for in terms of a relevance-based exhaustivity operator as well. However, the ordering relation needed to account for such examples can't be the topic-based ordering relation we have made use of until now in this section. In van Rooy & Schulz (ms) it is shown, however, that such an ordering relation can be induced by an *argumentative* oriented view of *relevance* as adopted by Ducrot (1973) and Merin (1998, 1999). In fact, it can be shown that both the purely cooperative topic-based ordering relation as defined by Groenendijk & Stokhof as well as an argumentative-based ordering between propositions can be seen to be special cases of a *utility theoretic* based ordering relation between propositions (cf. van Rooy, 2002). In different conversational circumstances – e.g. cooperative information exchange vs. argumentative discourse – a different instantiation of utility will be important, and thus a different ordering relation will be induced. This has an important consequence for our analysis of exhaustivity. The exhaustivity operator is based on an underlying ordering relation '>' that is always defined in terms of utility, but in different conversational situations the precise ordering involved might be different. We have made use of this assumption before – where the underlying question is either of the mention-all or the mention-some variety – but now we have come to an even more general perspective. Thus, with our new utility-based exhaustivity operator, we can now also account for scalar implicatures of the Hirschberg type.

5 Conclusion and Outlook

I have shown in this paper that there are major problems with the standard analysis of so-called 'quantity implicatures'. Especially the metalinguistic Q principle was seen to be problematic. I argued that most of these can be handled quite straightforwardly in terms of a relevance-based exhaustivity operator. This analysis differs in two crucial ways from the standard one: (i) the analysis is not metalinguistic, no alternative expressions are taken into account. The alternatives considered are all generated by the (generalized quantifier denoted by the) answer itself; (ii) making use of a relevance-based ordering relation means that we think of the invited inferences involved as crucially context-dependent.

The first contrast with the standard analysis suggests a way to solve the gap between the implicatures traditionally generated by the Q and I (Horn's R) principles. Indeed, we have shown already that some I implicatures (from *if* to *iff*) can be accounted for by means of exhaustification. But intuitively, at least, an inference to a stereotypical interpretation is very similar as well. Notice that our exhaustivity operator selects the *least relevant* interpretation(s) compatible with the meaning of the answer. As a special case, we take the *least surprising*, and thus *most likely* interpretation. But this special case comes down to the way Blutner (1998) implements the I principle to account for the stereotypical interpretation. This suggests that the relevance-bassed exhaustivity account is much more general than it appears at first sight.

Due to the second manner in which our analysis contrasts with the standard account – by taking *relevance* instead of *informativity* to be crucial – the inferences are thought of as *particularized* conversational implicatures rather than as *generalized* ones that can be *cancelled*. This doesn't mean that I think that there are no implicatures of the latter kind, nor that I don't see the context-independent notion of entailment playing a role. In fact, in van Rooy (to appear) I argue that Horn's division of pragmatic labor, or Grice's maxim of manner, should be thought of as a general *conventional* rather than a particularized *conversational* implicature. In other work I show that the importance of the context-independent notion of entailment can be given a utility-based explanation: for certain (cautious) decision procedures one can prove that entailment can be thought of as an *abstraction* from utility; A entails B iff receiving information A is *always* at least as useful as receiving information B. This suggests a reason why informativity-based implicatures have a greater chance to become conventionalized into semantics than other relevance-based ones: the first kind of inferences are relatively independent of the particular decision problem/question that is at issue.

References

[1] Atlas, J. and S. Levinson (1981), 'It-Clefts, Informativeness and Logical Form', In *Radical Pragmatics*, P. Cole (ed.), Academic Press, New York, 1-61.

[2] Bar-Hillel, Y. & R. Carnap (1952), 'An outline of a theory of semantic information', Technical report no. 247 of the *Research Laboratory for Electronics*, MIT.

[3] Blutner, R. (1998), 'Lexical pragmatics', *Journal of Semantics*, **15**, 115-162.

[4] Bonomi, A. and P. Casalegno (1993), '*Only*: association with focus in event semantics', *Natural Language Semantics*, **2**, 1-45.

[5] Carston, R. (1995), 'Quantity maxims and generalized implicature', *Lingua*, **96**, 213-244.

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

[7] Chierchia, G. (ms), *Scalar Implicatures, Polarity Phenomena, and the Syntax/Pragmatics Interface*, University of Milan.

[8] Ducrot, O. (1973), *La preuve et le dire*, Mame, Paris.

[9] Fauconnier, G. (1975), 'Pragmatic scales and logical structures', *Linguistic Inquiry*, **6**, 353-75.

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

[11] Grice, H. P. (1967), 'Logic and Conversation', typescript from the William James Lectures, Harvard University. Published in P. Grice (1989), *Studies in the Way of Worlds*, Harvard University Press, Cambridge Massachusetts, 22-40.

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

[13] Heim, I. (1992), 'Presupposition projection and the semantics of attitude verbs', *Journal of Semantics*, **9**, 183-221.

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

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

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

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

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

[19] Krifka, M. (1995), 'The semantics and pragmatics of polarity items', *Linguistic Analysis*, **25**, 209-258.

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

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

[22] Matsumoto, Y. (1995), 'The conversational condition on Horn scales', *Linguistics and Philosophy*, **18**, 21-60.

[23] McCawley, J. (1978), 'Conversational implicature and the lexicon', In *Syntax and semantics, Vol 9: Pragmatics*, P. Cole (ed.), Academic Press, New York, 245-259.

[24] Merin, A. (1999), 'Information, relevance, and social decisionmaking', *Logic, Language, and Computation, Vol. 2*, In L. Moss, J. Ginzburg, M. de Rijke (eds.), CSLI publications, Stanford, 179-221.

[25] Roberts, C. (1996), 'Information structure in discourse', In: J. Hak Yoon & A. Kathol (eds.), *Ohio State University Working Papers in Linguistics, volume 49.*

[26] Rooth, M. (1992), 'A theory of focus interpretation', *Natural Language Semantics*, **1**, 75-116.

[27] Rooy, R. van (2002), 'Utility, informativity, and protocols', In Bonanno et al (eds.), *Proceedings of LOFT 5: Logic and the Foundations of the Theory of Games and Decisions*, Torino.

[28] Rooy, R. van (to appear), 'Signalling games select Horn strategies', accepted for *Linguistics and Philosophy.*

[29] Rooy, R. van (ms), 'Questioning to resolve decision problems', University of Amsterdam.

[30] Rooy, R. van and K. Schulz (ms), 'Exhaustification', University of Amsterdam.

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

[32] Soames, S. (1982), 'How presuppositions are inherited: A solution to the projection problem', *Linguistic Inquiry*, **13**, 483-545.

[33] Sperber, D. & D. Wilson (1986), *Relevance; Communication and Cognition*, Blackwell, Oxford.

[34] Stechov, A. von and T.E. Zimmermann (1984), 'Term answers and contextual change', *Linguistics*, **22**, 3-40.

[35] Zeevat, H. (1994), 'Questions and Exhaustivity in Update Semantics', In: Bunt et al. (eds.), *Proceedings of the International Workshop on Computational Semantics*, Institute for Language Technology and Artificial Intelligence, Tilburg.