Informativeness, Relevance and Scalar Implicature*

Robyn Carston
University College London

1. Introduction

The main topic of this paper is the phenomenon of scalar implicature. Typical examples are given in (1)-(4):

(1) a. Bill has got some of Chomsky's papers.
b. The speaker believes that Bill hasn't got all of Chomsky's papers.

(2) a. There will be five of us for dinner tonight.
b. There won't be more than five of us for dinner tonight.

(3) a. X: I like Mary. She's intelligent and good-hearted.
Y: She's intelligent.
b. Y doesn't think Mary is good-hearted.

(4) a. She won't necessarily get the job.
b. She will possibly get the job.

The idea is that, in a wide range of contexts, utterances of the sentences in (a) in each case will communicate the assumption in (b) in each case (or something closely akin to it, there being a certain amount of contextually governed variation in the speaker's propositional attitude and so the scope of the negation). These scalar inferences are taken to be one kind of (generalized) conversational implicature. As is the case with pragmatic inference quite generally, these inferences are defeasible (cancellable), which distinguishes them from entailments, and they are nondetachable, which distinguishes them from conventional implicatures. The core idea is that the choice of a weaker element from a scale of elements ordered in terms of semantic strength (that is, numbers of entailments) tends to implicate that, as far as the speaker knows, none of the stronger elements in the scale holds in this instance. The pattern is quite clear in (1) and (2), where the weak/strong alternatives are *some/all* and *five/six* respectively. In the case of (3), the stronger expression must be *intelligent and good-hearted* which entails *intelligent*; what Y's utterance implicates is that Mary does not have the two properties: intelligence and good-heartedness, so that, given the proposition expressed (*Mary is intelligent*) it follows, deductively, that she is not good-hearted, in Y's opinion. The example in (4) involves a scale inversion due to the negation, so that the weak/strong alternatives are *not necessarily/not possibly*; the negation which the scalar inference generates creates a double negation, which is eliminated giving *possibly*.

Accounting for these sorts of examples, and more complicated scalar cases, has been, and still is, a central concern in neo-Gricean pragmatics (see references to Horn, Gazdar, Levinson, Hirschberg, Matsumoto, Welker, van Kuppevelt), but it has received relatively little attention in relevance-theoretic pragmatics. However, in the recent Postface to the second edition of *Relevance: Communication and Cognition*, Sperber & Wilson (1995) discuss instances of the

classic case in (1) above together with another well-known Gricean example, shown in (5):

(5) A: Where does C live?
B: Somewhere in the south of France.

The context is taken to be one in which it is clear that A wants a more precise answer (for instance, because she wants to go to visit C). The implicature Grice (1975) discusses here is one concerning B's inability to be more specific (because she doesn't know, or has forgotten, where exactly in the south of France C lives). Sperber & Wilson are also interested in a different possible implicature, one according to which the speaker is reluctant to disclose more specific information. In fact, as we will see, the apparent failure on the part of the speaker to say more (to be more specific, to make a stronger statement) in all five of the examples so far can give rise to either of these two distinct types of implicature, the "don't know" type or the "don't want to say" type, both of which, according to Sperber & Wilson (1995), follow directly from a reformulation of their fundamental communicative principle of relevance, and in a more satisfactory way than they do in Gricean frameworks.

My aims in this paper are modest: I bring together, and try to assess, a range of recent discussions of scalar implicature, conducted within several different post-Gricean[1] frameworks. My own contribution to the topic resides largely in an old, unpublished paper, Carston (1985), where I point out some of the distinguishing features of the numerical cases (such as (2) above), which show that a different treatment is called for, in both its semantic and pragmatic aspects, from that of the standard neo-Gricean accounts. I return to this matter in section 4.1 and discuss it further in the light of comments and criticisms by authors who have responded to it in the intervening years (in particular Horn 1992, Atlas 1992, van Kuppevelt 1996a, 1996b, Scharten 1997).

First, I will give some background to the current post-Gricean pragmatic scene, concentrating on the two criterial properties, informativeness and relevance, which have competed for centre stage in attempts to develop an inferential pragmatics governed by standards of communicative behaviour. Neo-Griceans find informativeness principles crucial to the explanation of scalar implicature; relevance theorists believe that the key concept in accounting for pragmatic inference, including scalar inference, is "optimal relevance".

2. Informativeness and relevance: a brief survey

2.1. *In the beginning*

The idea that utterances should meet some standard of informativeness could be seen as the starting point of current pragmatic theory. Grice (1961, 132) gives the following as his "first shot" at formulating a general principle governing the use of language: "One should not make a weaker statement rather than a stronger one unless there is a good reason for so doing". He suggests that something like this principle underlies the conversational implicature often carried by disjunctive utterances "P or Q". A statement of either "P (is true)" or "Q (is true)" would be stronger than the disjunctive statement so, given the general use principle, the speaker is taken to have implicated that she is ignorant of the truth-values of the disjuncts: she doesn't know P to be true and she doesn't know Q to be true. Somewhat earlier, in the context of a discussion of the divergences between logical operators and their natural language counterparts, Strawson (1952, 179, note 1) attributes to Grice the following pragmatic rule: "... one does not make the (logically) lesser, when

one could truthfully (and with greater or equal linguistic economy) make the greater claim". This version anticipates several of the maxims Grice suggested in his William James lectures in 1967, including his first maxim of quantity.

This well known collection of maxims, accompanying the overarching Cooperative Principle, includes two quantity (informativeness) maxims and a maxim of relevance. The first of the informativeness maxims says "make your contribution as informative as is required (for the current purposes of the exchange)" and the second "do not make your contribution more informative than is required". Grice expresses doubt about the second of these, as he considers it likely that a properly developed maxim of relevance would subsume it; he himself did not, to my knowledge, return to the issue of developing the concept of relevance. Versions of these two informativeness maxims have featured centrally in recent neo-Gricean pragmatics (especially that of Horn and Levinson). Strawson (1964), concerned with the issue of the topic or centre of interest of a statement, mentioned two "general platitudes" (his characterization) which he called "the Principle of Relevance" and "the Principle of the Presumption of Knowledge". The first of these is intended to capture the undoubted fact that "stating is not a gratuitous and random human activity. We do not, except in social desperation, direct isolated and unconnected pieces of information at each other, but on the contrary intend in general to give or add information about what is a matter of standing or current interest or concern." The second says that "statements, in respect of their informativeness, are not generally self-sufficient units, free of any reliance upon what the audience is assumed to know or to assume already, but commonly depend for their effect upon knowledge assumed to be already in the audience's possession" (Strawson 1964/71, 92). Despite their platitudinous nature and their apparent exclusion of utterances with purposes other than statement-making, these two principles are pointing in the right direction, I think. The requirements that utterance content connect up with existing assumptions and that the speaker take account of the hearer's current cognitive condition are met by the more inclusive and fully cognitively-grounded Communicative Principle of Relevance which is the heart of Sperber & Wilson's Relevance Theory. I leave detailed consideration of this account until section 5. The point that should emerge from the discussion to follow in the rest of section 2 is that no workable pragmatic system involving informativeness is able to function without drawing on considerations of relevance.

In one of the earliest assessments of the Gricean system, Harnish (1976 (in fact part of his thesis written in 1969)), after listing Grice's maxims says: "This maxim [relevance] turns out to be so central and important in conversational implicature that it is not clear that it belongs on equal footing with the rest. I suspect that maxims are (at least partially) ordered with respect to weight, etc. and that relevance is at the top, controlling most of the others." (Harnish 1976, 341, note 33).

Harnish was interested in establishing that the (a) sentences in the following examples do not entail the (b) sentences, but that utterances of (a) generally conversationally implicate (b):

(6) a. Russell wrote *Principia*.
b. Russell alone wrote *Principia*.

(7) a. The flag is red.
b. The flag is all red.

Like the examples in (1)-(4), these are cases of scalar implicature which are standardly treated as pragmatic inferences arising through some version or other of the first maxim of quantity. So in (7) the crucial part of the reasoning involves the recognition that a claim that the flag is red and

some other colour would be stronger than the claim made here, that it is red, (because "p and q" entails "p"), and therefore the speaker is implicating that the flag is not red *and* some other colour, and so that the flag is only red. In Harnish's system the work is done by his maxim of Quantity-Quality: "Make the strongest *relevant* claim justifiable by your evidence" (Harnish, 1976, 362, my emphasis), which effectively combines the second quality (truthfulness) maxim, the first quantity (informativeness) maxim and relevance. The advantage of a formulation that gives a central role to relevance is that it explains the absence of the *only*-implicature in particular contexts. For instance, say X and Y have a box of flags all of which are half blue and half white but which are distinguished by having a small block of some other colour in the white half: red, yellow or green. X is handing out flags to Y in some order or other and at a certain point Y is expecting to be given one that has a green patch on it. X makes a mistake and Y says:

(8) This flag is red.

Arguably, it would be more informative and equally well evidenced for him to say "This flag is blue, white and red", but it certainly would not be relevant. There are two points here: first, the example does not implicate that the flag in question is all red (since the totality of colour on the flag is irrelevant); second, it does implicate that the flag is not green, but it does not implicate that the flag is not blue (since the only *relevant* contrast set here is between a green patch and a patch of any other colour).

It is worth noting too that Harnish's quantity-quality maxim accounts for the implicature usually cited as arising for example (5) above, the "south of France" case: that the speaker is *unable* to be more precise about where C lives. She cannot make the relevant stronger statement because she doesn't have the knowledge/evidence to do so. However, neither this maxim nor any other in Harnish's system (or any of the other adaptations of Gricean maxims to follow) can account in any direct way for the alternative possible implicature mentioned above: the speaker is *unwilling* to give more precise information concerning C's whereabouts. More on these in section 5.

2.2. *Contrary forces: maximizing and minimizing informativeness*

Horn (1984, 1989) has developed an account which maintains Grice's Quality maxims (truthfulness and evidencedness)[2] but replaces all his other maxims with two general principles:

I. The *Q-principle*: Make your contribution sufficient; say as much as you can (given both Quality and R)
II. The *R-principle:* Make your contribution necessary; say no more than you must (given Q)

The Q-principle is taken to be a principle biassed in favour of the hearer's interest (to be given as fully articulated a verbal message as possible on the topic at hand) and is assumed to encompass Grice's first maxim of Quantity (Make your contribution as informative as is required) and to mop up the first two Manner maxims ("Avoid obscurity of expression" and "Avoid ambiguity"). The R-principle, on the other hand, is taken to be a principle biased in favour of the speaker's interest (to expend as little articulatory [and cognitive] effort as possible) and is assumed to subsume Grice's second maxim of Quantity ("Do not make your contribution more informative than is required"), his maxim of Relation and the other two Manner maxims ("Be brief" and "Be orderly") (see Horn 1989, 194).

So he sees these principles as pulling in opposite directions and as reflections within the

sphere of communication of deeper contradictory forces at work in language change: Zipf's principle of least effort (speaker's economy), on the one hand, which taken to its logical extreme would result in a single vocable encoding all meanings, and his "force of diversification" (hearer's economy), on the other hand, which taken to its logical extreme would result in a vast vocabulary of distinct words, one for each meaning. Whatever the value of this view of things for language change, it strikes me as quite wide of the mark when it comes to verbal communication. Speakers are often very interested in being understood, having their message received, and this must, at the very least, modify their alleged concern to keep their articulatory organs in repose. The assumption that hearers would really like to have every ounce of intended meaning enshrined in linguistic form is equally dubious; the psychological evidence indicates that our cognitive systems are finely attuned to aspects of context, including states of mind of speakers, so that explicit encoding, over a fairly low threshold, is more likely to impede than enhance communication. Once a few basic facts concerning human cognitive processing of proximal stimuli generally, such as the massive (virtually automatic) inferential contribution of the system to an often meagre input, are taken on board, and once we set our account of utterance understanding in this cognitive context, Horn's assumptions about speaker's and hearer's best interests seem straightforwardly wrong. Relevance theory is such a cognitively based account; it sees a speaker as having a primary interest in getting her message across and a hearer as following a comprehension strategy of keeping his processing effort to a minimum which entails not being giving a lot of linguistic material to decode when its content is already activated or readily inferable. This is developed further in section 5.

It is clear, though, from the description of the effect of these two principles that they may "clash", and Horn (1984, 19) gives the following examples as evidence of the opposing sorts of implicature they can give rise to:

(9) I slept on a boat yesterday.
implicates: The boat was not mine.

(10) I lost a book yesterday.
implicates: The book was mine.

The implicature in (9) is a result of the Q-principle and is a typical scalar implicature, *my* being higher on a particular entailment scale than the indefinite article. The implicature in (10) is a result of the R-principle; it is a case where the speaker is taken to have spoken minimally and so to mean more than she said, so that a "stereotypical relation" is assumed to hold between the book and the speaker. Several critical points could be raised here concerning such issues as what determines which of these principles, with their opposing results, comes into force when, how the appropriate scale of items is determined in the Q cases, and how the hearer knows how much informational enrichment to supply in the R cases. These issues arise with equal urgency in the context of Levinson's work to be discussed next.

A point worth noting before moving on is that, as Green (1995, 89) points out, the two formulations of each of the principles (one on each side of the semi-colon) are not equivalent. For instance, a conversational contribution which is informationally sufficient need not involve the speaker in making the strongest statement she is able to make on the matter at hand (i.e. saying as much as she can). With regard to the Q-principle, Green shows that of the two versions here the second is stronger than and entails the first; he argues that the second, which he dubs the *principle of volubility*, is a widespread misinterpretation of Grice's first quantity maxim (Grice 1967/89) and gives rise to some false predictions. Some of the earliest attempts at formulating a pragmatic

principle (for instance, Grice 1961; see section 2.1 above) came up with volubility (rather than informational adequacy), and, less forgivably, so did various later interpretations of Grice's 1967 system: Harnish (1976)'s quantity/quality maxim given above, Levinson (1983, 106), Hirschberg (1985/91) and Matsumoto (1995). Some of Green's examples of wrong predictions by the neo-Griceans' volubility principle are considered in section 5, where the predictions of Grice's first quantity maxim, the volubility principle and relevance theory are compared.

The two main principles (with their accompanying inferential heuristics) in the system of Levinson (1987a, 1988) are very similar to, though different in detail from, Horn's. These are the Q-principle and the I-principle, which seem to pull in opposite directions. His Q-principle is more or less identical to Horn's, while his I-principle, a principle of informational enrichment, is similar to Horn's R-principle but cannot, in his view, subsume the manner maxims or the relevance maxim. He formulates a distinct M-principle[3] but offers nothing by way of characterisation of a relevance principle. That it is essential and distinct from the quantity principles in his system is evident in his account of the hearer's application of the I-principle (which gives rise to implicatures which are strengthenings or enrichments of the content of what is said): "amplify the informational content of the speaker's utterance, by finding a more *specific* interpretation, up to what you judge to be the speaker's m-intended point" (Levinson 1987a, 68). How does one judge what the speaker's m-intended point is? The answer seems to be via considerations of relevance. Given that in any application of the Q-principle the appropriate expression alternatives, or scale, also depend on relevance, it looks as if, in this picture, it is relevance that controls and orders the application of other principles/maxims in the system, as Harnish (1976) overtly acknowledged. These points apply equally to Horn's system, despite his claim that his R-principle subsumes relevance.

I have already addressed Levinson's ideas in some detail in Carston (1990/95), so will focus in the next section on Horn's principles, though much of what I have to say applies to both of them. Ideas which they hold in common include at least the following:

[1] There are two central "informativeness" principles that apparently pull in opposite directions: the one principle seems to enjoin maximal informativeness, the other minimal, and they each give rise to a class of implicatures which are opposed, in that the first involve negations of stronger propositions while the second are strengthenings of what was literally expressed.

[2] Conversational implicatures fall into two classes: generalized and particularized, and this is a theoretically interesting distinction. This is much more pronounced in Levinson's work than Horn's. Levinson has developed a distinct theory of **generalized** conversational implicatures, according to which they are generated by a set of default inference rules (see Levinson forthcoming).[4]

[3] A central and apparently unified subclass of generalized implicatures is the class of scalar implicatures, dependent for their derivation on the Q-principle. This has been a main focus of Horn's work since his 1972 thesis and has received considerable subsequent attention from others working in various modified Gricean frameworks. Part of its appeal is that it has seemed more amenable to formal, including computational, modelling than any other type of conversational implicature (see, in particular, Gazdar 1979, Hirschberg 1985/91, Wainer & Maida 1990, Welker 1994, Iwanska 1996).

Before looking more closely at the Horn/Levinson programme I'll finish this section with a brief discussion of some other post-Gricean treatments of scalar implicatures and of the role that relevance plays in their accounts.

2.3. *Informativeness rankings and contextual relevance*

Hirschberg (1985/91)'s work on scalar implicature marks a considerable widening and deepening of formal and computational work in the area, while maintaining an essentially Gricean framework (following Horn (1972) and Gazdar (1979)). Leaving aside here the formal aspects of her account, the following features are significant:

[A] Horn's linearly ordered scales of items in an entailment relation (quantifiers, modals, numerals, logical connectives) are but one small subset of the sorts of orderings that give rise to implicatures via a reasoning process driven by the first maxim of quantity. Various other linear and hierarchical relations which impose a *partial ordering on sets of elements* work in the same way. These include rankings of entities, states, and attributes; whole/part relationships; type/subtype, instance-of (isa), and generalization/specialization relations; entity/attribute relations. The following examples are mostly taken from Hirschberg:

(11) ranked entities:
A: Is Jill a professor yet?
B: She's a senior lecturer.
implicature: Jill isn't a professor

(12) whole/part relation:
A: Did you manage to read that chapter I gave you?
B: I read the first couple of pages.
implicature: B did not read the chapter.

(13) instance-of:
A: Do you have any juice?
B: I have grape, orange and tomato.
implicature: B doesn't have any lemon/apple/etc.

[B] Inferences may be based on an alternate value (rather than a higher one) in a non-linear ordering:

(14) A: Did you get Paul Newman's autograph?
B: I got Joanne Woodward's.
implicature: B didn't get Paul Newman's autograph.

(The assumption here is that Paul Newman and Joanne Woodward are ranked at the same level, like the various juices in the next example; another possibility, of course, is that in this context Paul Newman is ranked higher than Joanne Woodward on a linear scale of famous movie stars.)

(15) A: Do you have apple juice?
B: I have grape or tomato or orange.
implicature: B doesn't have any apple juice.

[C] Inferences may arise not just from a speaker's affirmation of some value but also from a speaker's denial of or assertion of ignorance of some value. Denial of a value implicates the truth of alternate values:

(16) A: Have you made fondue in this pot yet?

B: Not chocolate fondue.
implicature: B has made some kind of fondue in the pot.

Assertion of ignorance of some value implicates the truth of a value lower on the scale and the falsity or unknownness of a value higher on the scale:

(17) A: Do you have information on [Kathy M. for maternity] ...?
B: I don't think she has delivered yet.
A: Then she HAS been admitted.
B: Yes.

B's first utterance implicates that she believes that Kathy M. has been admitted (*admission* being a lower value than *delivery* on the scale of activities in the process of having a baby), an implicature which A's second utterance checks on.

Hirschberg's account is more context-sensitive than previous formal attempts, so that, for instance, context plays an important role in determining which implicatures will arise, rather than merely having the role of cancelling previously generated potential implicatures (as in Gazdar (1979)'s formal treatment). Furthermore, while she runs her account on a combination of the first maxim of quantity and the maxims of quality, she acknowledges that there is a crucial, yet to be specified, role for a properly defined notion of relevance in, for instance, determining the type of ordering at issue in a particular context and the specific elements within that ordering (p.65). Her account does not cover either of the possible implicatures of the "south of France" example discussed above.

Welker (1994, 31) makes an interesting point about the parenthetical comment in Grice (1967/89)'s formulation of the first maxim of Quantity, "... as informative as is required (*for the current purposes of the exchange*)"; she says that this indicates that the operation of Quantity 1 is "bounded" by Relevance, so that "the Maxim of Relevance may be more important than Quantity 1". She goes on to discuss the fact that while a scalar implicature ("at most two") arises in (18), it tends not to in examples like (19) (which is not understood as communicating "at most four"):

(18) A: I'm having a dinner party and I need four more chairs.
B: John has two chairs.

(19) A: I'm having a dinner party and I need four more chairs.
B: John has four chairs.

She follows the standard view that a cardinal number scale is involved in both cases, a scale which usually makes available a range of semantically stronger expressions and so enables a comparison between what was actually said and certain similar but more informative propositions which could have been expressed. Following Hirschberg (1985/91), Welker assumes that the reason for the difference is that, in both examples, the highest point on the numerical scale established in the context, as a result of A's utterance, is *four*, since possession of a higher numbers of chairs is irrelevant to A's concerns. As B's utterance in (19) is maximally informative relative to this scale, no implicature arises. Relevance prevails over Quantity 1 in that it plays the fundamental role of establishing both the type of scale and the end-points of the scale over which Quantity 1 operates. Roberts (1996) reiterates and reinforces this point.

A similar point can be made with regard to another set of examples for which the volubility principle, unconstrained by relevance considerations, makes the wrong prediction:

(20) A: What did you buy for your mother?
B: I bought her flowers.
predicted scalar implicature: I didn't buy her roses.

(21) Billy got a dog for Christmas.
predicted scalar implicature: Billy didn't get a spaniel.

These could be seen as involving either a Horn entailment scale (e.g. <rambler, rose, flower, plant> or an "instance-of" partial ordering of the sort Hirschberg endorses. Here the question is: what prevents these quantity-based implicatures from arising? There is a lot of psychological evidence around that in taxonomies of biological terms there is a "basic level" category which is the unmarked level of reference; *flower* and *dog* are the terms for the basic level categories of their taxonomies. Hirschberg (1985/91, 160) says that the fact that generally, in the absence of special conditions, a scalar implicature does not arise when these are used, is due to the relevant scales being upper-bounded by the basic level term. This is a natural, though not very explanatory, move to make in a volubility-based framework. Alternatively, and perhaps more explanatorily, we could say that some pragmatic criterion other than that proposed by Horn, Harnish, Levinson, and Hirschberg is in operation here: although the speaker could have made a stronger statement, one which specifies the sort of flowers or the sort of dog, the informationally weaker statement is *sufficiently relevant* (that is, it has enough cognitive effects) to be worth the hearer's attention and so does not give rise to any implicature that the stronger proposition is not the case. (This sort of case is also discussed by Matsumoto (1995, 29) and Scharten (1997, 37).)

Matsumoto (1995), a strong supporter of a neo-Gricean account of scalar implicatures, develops an account of what constraints the first maxim of quantity operates under. Following Horn and Hirschberg, his formulation of quantity-1 is a volubility principle, rather than Grice (1967/89)'s informational adequacy maxim. He claims that a scale <S, W> (where S is the stronger item and W the weaker) licenses a scalar (quantity-1) implicature only if the following condition is met:

> *Conversational condition:* the choice of W instead of S must not be attributable to the observance of the maxims of quantity-2, relation or obscurity avoidance (manner-1).

This is a (somewhat convoluted) way of saying that observing the quantity-2, relevance and non-obscurity maxims takes precedence over observing quantity-1 (= volubility, on this account). As has already been remarked, quantity-2 will be subsumed by an adequate account of the relevance principle. The Sperber-Wilson characterization of relevance also subsumes and finesses the requirement to avoid obscure expressions; such expressions are to be avoided when they cause unjustifiable processing effort, that is, effort that is not offset by extra cognitive effects. So, once again, it looks as if relevance is the key constraint. Matsumoto (1995) recognises this possibility, but argues that there are cases for which the Sperber-Wilson account of relevance makes the wrong predictions; this is discussed in section 5.3.

The main point of this second section of the paper has been to show that even those pragmatists most committed to principles concerning quantity of information have no choice but to advert to considerations of relevance, even when they are explaining the paradigm case of quantity-based implicatures, namely scalar implicatures. They all rely on an intuitive notion of relevance, none having attempted any precise characterization of the notion or of the factors that go into making an utterance more or less relevant. So an obvious question is whether a properly formulated account of relevance might not do the job without the need for any independent

informativeness principle. This question is addressed in section 5, where I try to show that the single communicative principle of relevance that emerges from the more general characterization of cognitive relevance developed by Sperber & Wilson (1986/95), does account for the full range of implicatures, making redundant any principles or maxims which enjoin speakers to be minimally, adequately or maximally informative.

I return in the next section to what is still, without doubt, the dominant post-Gricean account of scalar (and other quantity-based) implicatures, that associated with Horn and Levinson; I take a critical look here at the way in which their quantity principles have been put to work and at the allegedly contrary pragmatic inferences they give rise to.

3. Quantity implicatures

3.1. *Applications of the Q-principle and I/R principle*

In this short section I will point out the considerable work that these two quantity principles, though more particularly the Q-principle and its resultant Q-implicatures, have been put to in accounting for a wide range of phenomena. For instance, Levinson (1987b, 1991) has proposed accounts of coreference and disjoint reference which employ scalar implicatures as part of the machinery which determines the bindings speaker/hearers find acceptable; this is argued to be more adequate than a syntactic account in terms of Chomskyan Binding Conditions. One element of these accounts is a Horn scale consisting of the abstract grammatical categories "reflexive" and "pronoun", where "reflexive" is the stronger of the two, since it is necessarily referentially dependent, while the pronoun is only optionally so. Given various locality conditions, choice of the weaker pronoun (where a reflexive could have been used) is taken to give rise to a scalar implicature to the effect that the reference (within a given domain) is disjoint (that is, not coreferential). In other contexts where the <reflexive, pronoun> scale does not come into play, the use of a pronoun implicates local coreference via the other quantity principle (I or R), which typically enriches the intrinsic semantics. In similar vein, Gundel et al. (1990, 1993) use the first quantity maxim to account for which referring expression a speaker chooses from a hierarchy (or scale) of givenness and thus for what a hearer takes her to be communicating on the basis of this choice. For example, demonstrative pronouns (*this*, *that*) indicate that the referent is activated, where "activated" has a lower status on the givenness hierarchy than "in focus", which is indicated by a personal pronoun (for instance, *it*); it follows, then, that a speaker who chooses to refer to a particular entity by using *this* rather than *it* communicates a scalar implicature to the effect that the referent is not in focus.

Another example of this use of scalar implicature as a kind of processing instruction to a hearer, constraining the pragmatic inferential phase of comprehension, is Oberlander & Knott (1996)'s account of what they call cue phrases. These include discourse connectives such as *so*, *but*, *therefore*, and *whereas*, which are standardly assumed to fall outside the representational (truth-conditional) content of the proposition expressed by an utterance[5], but also include cases that are plainly truth-conditional, such as *as soon as* and *as a result*. The claim is that these divide up into partially ordered sets of elements to which the first quantity maxim applies. For example, *whereas* and *despite this* are hyponyms of *but*, so a speaker who chooses *but* rather than one of the more specific cue phrases implicates that the stronger relation (between the conjoined clauses) does not hold. As the authors point out, an alternative possibility is that the speaker/writer uses the less specific form when she can rely on the hearer's context to enrich to the more specific relation; that is, a strengthening is effected by the second maxim of quantity (Levinson's I-principle

or Horn's R-principle).

I am extremely dubious about aspects of these approaches: they seem to make the term "implicature" incoherent; it is used indifferently to apply to both implicitly communicated assumptions and to some of the processing indications provided by a speaker to help hearers arrive at the intended assumptions. A speaker who uses the pronoun *it* does not communicatively intend to inform her hearer that an entity she is referring to is not in the focus of his attention and a hearer doesn't entertain (or later recall) this as part of what the speaker communicated to him. Similarly, a speaker who utters *Bob was unwell; but he gave a great lecture*, rather than *Bob was unwell; despite this he gave a great lecture*, does not communicatively intend to inform the hearer that Bob's giving of a great lecture does not violate an expectation created by his unwellness. There is a conflation here between *how* certain communicative tools (pronouns, cue phrases) work and *what* is communicated by using them.

At a more fundamental level still, there is reason to doubt that there really are two such principles making opposing recommendations regarding quantity of information and giving rise to two different types of quantity implicature, as envisaged by Horn and Levinson; I turn to this point in the next section.

3.2. *Two types of generalized quantity implicature?*

I repeat the first three examples of scalar implicatures given in the introduction:

(1) Bill has got some of Chomsky's papers.
implicating: Bill hasn't got all of Chomsky's papers.

(2) There will be five of us for dinner tonight.
implicating: There won't be six/seven ... of us for dinner.

(3) A: I like Mary. She's intelligent and good-hearted.
B: She's intelligent.
implicating: She's not good-hearted.

It is usually noted that these implicatures involve the negation of a proposition which is semantically stronger than the proposition expressed by the utterance itself. This type of implicature is compared by Atlas & Levinson (1981), Horn (1984, 1989) and Levinson (1987a) with what is alleged to be the opposite sort of case, where what is implicated is straightforwardly stronger than the proposition expressed (i.e. negation does not enter into the picture):

(22) He drank a bottle of vodka and fell into a stupor.
implicating: His falling into a stupor was a result of his drinking a bottle of vodka.

(23) Sam and Mike moved the piano.
implicating: Sam and Mike moved the piano together.

(24) If you finish your thesis by September you'll be eligible for the job.
implicating: You'll be eligible for the job if and only if you finish your thesis by September.

Notice that the implicature here entails the proposition expressed by the utterance, that is, it is

semantically stronger; these are said to be the result of the countervailing pragmatic maxim, the I-principle (Atlas & Levinson 1981, Levinson 1987a) or R-principle (Horn 1984, 1989), a principle which gives rise to implicatures which enrich or strengthen the proposition literally expressed.[6] Much work has been done in trying to devise a system which accounts for when the Q-principle comes into operation and when the R-principle does. This has seemed essential since they appear to be making opposite predictions, so that the R-principle applied to (1) would give the wrong result ("Bill has all of Chomsky's papers") and the Q-principle applied to (22) would give the wrong result ("His falling into a stupor was not the result of his drinking a bottle of vodka"). Some of this work has been directing at placing conditions on what constitutes a valid scale (<all, some> is well-formed while <P and as a result Q, P and Q> is not) and some at placing contextual constraints on when the Q-principle comes into operation (see Atlas & Levinson (1981), Levinson (1987a), Horn (1989, section 4.4) and, in particular, Matsumoto (1995)).

The idea behind the informational enrichment principle is that speakers should leave unsaid that which is obvious, noncontroversial or stereotypical, since the hearer can easily supply this himself. So, for instance, the temporal and cause-consequence relations that are often pragmatically inferred in *and*-conjunction utterances, as in (22) above, are stereotypical and will be taken by the hearer to be intended by the speaker in the absence of indications to the contrary.[7]

However, consider the following two examples, the first of which is an I-implicature (from Atlas & Levinson (1981), 41) and the second of which is a Q-implicature (of my own contriving):

(25) John was reading a book.
implicates: John was reading a non-dictionary.

(26) Some people like eating raw liver.
implicates: Not everyone likes eating raw liver.

Atlas & Levinson's line of reasoning for (25) is that in stereotypical situations involving the reading of a book, the book is not a dictionary, so the hearer is licensed to derive the informationally enriching implicature that narrows down the non-specific predicate "book" by excluding dictionaries. Example (26), on the other hand, is an example of the standard scalar implicature type involving the scale <all, some>. Given the system of pragmatic principles taken to be at work here, these explanations are certainly available, but, as far as I can see, each of these examples might just as readily be given the opposite sort of explanation. So the negative implicature in (25) could be the result of the speaker's having chosen the lower item from the scale <dictionary, book>. And the implicature of (26) could be viewed as providing an informational enrichment of the proposition literally expressed, one whose content is obvious/stereotypical (raw liver is generally assumed to be unappetising) in the way that is typical of I/R-implicatures. In these cases, at least, the principles do not seem to be clashing at all but rather converging on the same result, the result in both instances being an enrichment or narrowing down of the options left open by the proposition expressed. This might lead us to wonder if there is not some deeper, more all-encompassing, principle at work here which subsumes these two seemingly distinct maxims.[8]

In fact, the alleged clash of principles, and consequently different types of implicature they give rise to, has already been shown to be more apparent than actual (see Carston 1990/95; Richardson & Richardson 1990, (hereafter R&R)). It seems to be based on a confusion between what is implicated and what is conveyed/communicated by an utterance (which includes what is implicated). Following R&R, let's take a look at the supposedly opposing pragmatic schemes, given in (27), where "S" is the stronger and "W" the weaker term and (to keep things simple) relative strength is measured in terms of numbers of entailments. An example of each follows in

(28):

(27) *Q-based implicature:*
S entails W
"W" implicates "not S"

R-based implicature:
S entails W
"W" implicates "S"

(28) a. [P and Q] entails [P or Q]
"[P or Q]" Q-implicates "[not [P and Q]]"

b. [P iff Q] entails [if P, Q]
"[if P, Q]" R-implicates "[P iff Q]"

But, as I have shown (Carston 1990/95, 219-220), there is an odd disparity in the way these examples are described; if we observe the three-way distinction between what is said, what is implicated and what is conveyed (which includes both what is said and what is implicated), as we surely must, we get the following:

(29) a. what is said: P or Q
what is implicated: not [P and Q]
what is conveyed: P or Q but not both
or: just one of P, Q

b. what is said: if P, Q
what is implicated: if not P, not Q
what is conveyed: P iff Q

where, in both cases, what is conveyed entails what is said and what is implicated is logically independent of what is said (not in an entailment relationship with it). In short, both involve a pragmatically derived strengthening of what is said, contrary to what the schemes in (27) indicate. **All** the above examples, whether Q-based or R/I-based, involve informational enrichment of what is said; as R&R (1990, 507) put it: "What Horn, Atlas and Levinson have lost sight of is that it is quite generally the case that what is conveyed properly includes and therefore entails what is said".[9]

R&R take this a step further, showing that the two examples could just as easily be given as cases of the opposite sort of implicature to the one they are taken by Horn and the others to exemplify. So the exclusive understanding of "P or Q", which entails the inclusive understanding could be seen as a case of R-implicature and, given a bit of playing about with logical equivalences, the conditional could be claimed to generate a Q-implicature (that is, the negation of a stronger proposition):

(30) a. [P or-excl Q] entails [P or-incl Q]
"[P or-incl Q"] R-implicates [P or-excl Q]

b. [not [if(not P), not Q]] entails [if P, Q]
"[if P, Q]" Q-implicates [if(not P), (not Q)]
(by double negation)[10]

They complete their devastating criticism with the following two examples, the first of which

Grice (1975) used to illustrate generalized conversational implicature, and which Horn has taken to be a case of R-implicature, as opposed to other cases of the indefinite article which are supposed to give rise to Q-implicatures (and still others that do neither). R&R's example of the opposite type is given in (31b) and demonstrates very vividly that either of the opposing schemes in (27) is as good (or poor) as the other in accounting for both examples and that there is no reason to prefer one to the other.

(31) a. I broke a finger.
implicates: I broke one of my own fingers.

b. I found a finger.
implicates: I found someone else's finger.

In both cases pragmatic inference results in a proposition which is stronger than (entails) the ownership-neutral linguistic content of the utterance. Of course, the inferences result in an opposition of sorts, "my" finger versus "someone else's", but this is patently a consequence of bringing to bear general knowledge assumptions on a quite general pragmatic process of strengthening, arguably driven by a single maxim or principle of adequate informativeness/relevance. I endorse R&R's conclusion that this account in terms of allegedly opposing informativeness maxims cannot succeed "until some mechanism is proposed for predicting which entailing propositions are *relevant* for predicting the implicatures of an uttered proposition and which are not" (R&R 1990, 507, my italics).

R&R's paper is pitched against the "radical pragmatics" programme quite generally, which includes the relevance-theoretic approach. They mention the theory in passing, dismissively, and clearly assume that it too will fall prey to the same (and probably further) criticisms since it is even more minimalist than the Horn/Levinson approach, having but "one grand principle". I hope to show (in section 5 below) that this is not in fact the case, and that the cognitively-based pragmatic criterion provides just the guidance that is needed to account for, among other cases, the different strengthenings of the examples in (31). It should be noted that while R&R's critique is penetrating and must be addressed, they themselves offer absolutely nothing by way of a positive account of the pragmatic phenomena they present in their paper.

From this point on, I shall leave aside the pragmatic inferences standardly listed under the I/R label and concentrate entirely on those that have been given a Q analysis, since these have not received as much attention within the relevance-theoretic framework. Some, at least, of the cases that the neo-Griceans treat as I/R-implicatures have received detailed analysis within relevance theory (see, for instance, Carston 1988, 1990/95, 1993, 1994; Wilson & Sperber 1993b (and in this volume)). The bare essentials of our account of the informational enrichments of *and*-conjunctions are the following: they are not implicatures at all but rather constitute a pragmatic contribution to the proposition expressed by the utterance, hence to its truth-conditions; their occurrence is accounted for in terms of highly accessible general knowledge schemas, some general facts about the processing of sequentially presented information and the constraints imposed by the search for an interpretation consistent with the communicative principle of relevance.

4. Semantics and pragmatics of quantitative terms

I shall largely avoid the term "scalar implicature" in this section, since the sort of pragmatic inference we are looking at does not necessarily involve a *scale* of elements and rather than

eventuating in an *implicature*, it may contribute to the proposition expressed by the utterance and so to its truth-conditions.

4.1. *Pragmatic inference: the number terms versus the rest*

Horn (1972, 1989 (chapter 4)) argues that cardinal numbers ought to be given an "at least" semantics, so that *I have five pounds* has the same truth conditions as *I have at least five pounds*, although an utterance of the former often implicates that the speaker has no more than five dollars, so that what is communicated is that the speaker has "exactly" or "only" five pounds ("at least five" together with "at most five" gives "exactly five"). Most other writers on scalar inferences issuing from cardinal number terms have made the same assumption about what the cardinals encode and how the narrower "exactly" understanding is derived (Gazdar 1979, Levinson 1983, 1987a, Hirschberg 1985/91). The account follows the general scheme for the bilateral understanding of scalar terms (for example, *some*, *possible*, *or*), according to which an (upper-bounding) scalar implicature (for example, *not all*, *not necessary*, *not both*) is added to a lower-bounding semantics.

However, the following neat point made, more or less in passing, by Harnish (1976, 326) should have put paid to this idea for number terms long ago:

> ... suppose you bet me that there will be 20 people at the talk tonight. We arrive and there are 25 people there. Who wins? There may be some temptation in both directions, but that seems to be because the question is underdetermined. It seems that the sentence
>
> (32) There will be 20 people there.
>
> can be used to make the following claims:
>
> (33) a. There will be at most 20 people there.
> b. There will be exactly 20 people there.
> c. There will be at least 20 people there.
>
> Of course I do *not* want to claim that (32) is ambiguous and has (33) as its senses. Suppose that in the situation imagined, I had been complaining about the poor attendance at talks and you reply with (32) - against the mutual understanding that 20 people is a good turnout. In this context, what you said could have been paraphrased as (33c), and so you would win the bet. Another context could have changed the force of my utterance to either of the other two.

This highlights two problems with the neo-Gricean account: (a) it ignores the "at most" understanding given in (33a) above and, given its favoured semantics, it cannot account for it; (b) whichever of (33a)-(33c) is the correct interpretation on a given occasion, that interpretation is taken to be explicitly communicated, in fact to constitute the truth-conditions of the utterance. I have argued these points before (Carston 1985, 1988, 1990/95); here I will briefly survey other work on these issues.

Let's start with the second point, as it seems now to be quite widely accepted. If the proposition expressed on a given occasion does involve the "exactly n" understanding of a number term, then it cannot be the case that the upper-bounding pragmatic inference constitutes an implicature, since implicated propositions are distinct from the proposition expressed by the

utterance and cannot affect truth-conditions. Rather, this would be a case of a pragmatic inference which plays the role of enriching an underspecified logical form; there are many such occurrences of pragmatic inference due to the considerable underdetermination of the proposition expressed by encoded linguistic meaning. The general point has been argued widely (see Travis 1985, Sperber & Wilson 1986/95, Kempson 1986, Carston 1988 and forthcoming, Recanati 1989, Atlas 1989) and has been alluded to in this paper in the brief discussion of the *and*-conjunction cases. As regards the specific case of the cardinal number terms, several authors have presented additional arguments in recent years for this being the right way of viewing the upper-bounding pragmatic inference they often give rise to, and for thereby distinguishing them from the other scalar terms, which still seem appropriately accounted for by the traditional implicature account.

Sadock (1984) pointed out that the Hornian semantic-pragmatic account of the number terms does not seem able to account for examples such as those in (34a) and (34b); for instance, (34b) would come out as true on the "at least" understanding of the cardinals. Richardson & Richardson (1990, 501), in a general critique of the "semantic minimalism" of the radical pragmatics programme, endorse Sadock's point and, lest his example be dismissed as a technical, mathematical use of the number terms which intrinsically requires an "exactly" understanding, they give example (34c), which is "a perfectly colloquial and clearly quantificational sentence beyond the semantic reach of Horn [1972, 1989]'s proposal":

(34) a. The square root of nine is three.
b. Two plus two is three.
c. I took six cigarettes with me, gave one to Fred and two to Ed, so I still have three.

In other words, the account on which the "at most" element of the communicated meaning of the cardinals is treated as an implicature makes the wrong predictions about truth-conditions. (Although Sadock and R&R are making the same criticism of the neo-Gricean account, they differ considerably in their positive proposals regarding the correct semantics of the cardinals, as we'll see in the next section.)

Horn (1992) makes a number of observations in support of a pragmatic enrichment account of the interpretation of cardinals. In addition, he shows that this does not carry over to other scalar cases. For instance, he contrasts the following two examples:

(35) A: Do you have two children?
B: No, three.
C: ? Yes, (in fact) three.

(36) A: Are many of your friends linguists?
B: ? No, all of them are.
C: Yes, (in fact) all of them are.

In (35B), the perfect compatibility of *three* with the negative answer indicates that the negation is taken to be denying the having of exactly/just/only two children rather than the having of at least two. The corresponding response in (36B) is marked; it CAN be processed in a way that makes it a consistent response, but only after a kind of reanalysis prompted by the follow-up clause, *all of them are*, and the reanalysis seems to involve what's known as metalinguistic negation. For Horn (1985, 1989) the negation is being used to object to a scalar implicature (*not all*) of the question; for me, there is a tacit echoic use of *many* falling within the scope of the negation

(Carston 1996). The naturalness/markedness judgement is the reverse for the affirmative responses in (35C) and (36C), again indicating that the pragmatic inference affects truth-conditions in the number case but not in the other. Horn further remarks that a bare *No* answer is compatible with an "exactly two" understanding in (35) [given the fact that B has three children], while this is never the case in (36), as a bare *No* can only be understood as conveying "less than many".

Horn (1996, 316) reinforces the point with further examples:

(37) a. ??Neither of us liked the movie - she hated it and I absolutely loved it.
b. Neither of us have three kids - she has two and I have four.

Negation of the non-cardinal scalar term, *like*, in (37a) cannot be interpreted as a denial of the enriched two-sided content, while this is fine for the cardinal term.[11] He concludes: "Such paradigms support a mixed theory in which sentences with cardinals may well demand a pragmatic enrichment analysis of what is said, while other scalar predications continue to submit happily to a minimalist treatment on which they are lower-bounded by their literal content and upper-bounded, in default contexts, by quantity implicature." (Horn 1996, 316).

A further pointer in this direction comes from Scharten (1997, 67-68), who notes an erroneous asymmetry in the implicature view: in (38), B's utterance is considered a case of implicature cancellation, while in (39) it is a case of repair or self-correction:

(38) A: How many pupils are there in your class?
B: 31. No wait, 33.

(39) A: How many pupils are there in your class?
B: 31. No wait, 29.

Intuitively, the two cases are exactly parallel, and this is explained on an account according to which the truth-conditional content of B's first utterance in each case is that there are exactly 31 pupils in B's class, and the follow-up clause in both cases is a self-correction. Scharten develops an account along these lines, one which she believes is not restricted to the number terms but extends to the full range of cases that Horn and Hirschberg treat as supporting scalar implicature.

She follows van Kuppevelt (1996a, 1996b) in claiming that the crucial factor determining whether a cardinal (or other scalar term) is given an "at least" or an "exactly" interpretation is whether the term is in the topic or the comment part of the information structure of the utterance. In brief (and omitting certain subtleties), if it occurs in a topic it gets an "at least" interpretation and if it occurs in a comment it gets an "exactly" interpretation.[12] This work is done in the framework developed by van Kuppevelt (1991, 1995) called Discourse Topic Theory (DTT), according to which discourse is essentially structured around topics (and subtopics), established by questions (many of which are implicit), and a well-formed unit of discourse is one in which the topic-establishing question has been satisfactorily answered (the comment has supplied the required value for the indeterminacy in the question). There is no role for a quantity (informativeness) maxim or any other pragmatic principle in this account. Simple examples involving cardinal number terms are the following:

(40) How many children does John have?
a. He has **three** children.
b. * He has **three** children, in fact **five**.

(41) Who has three children?

a. **John** has three children.

b. **John** has three children, in fact he has five.

(where the comment is in bold)

In (40), where the numeral occurs in comment position, there is an "exactly" reading, so that the addition of *in fact five* makes the utterance incoherent. In (41), where the numeral is in topic position (the topic established by the question being "the having of three children"), the understanding is merely lower-bounded, so that the addition of *in fact five* is fine. In this example, it is the name *John*, which is in comment position, which induces a scalar inference: a denial that any other candidates in the given context have (at least) three children.[13]

Van Kuppevelt (1996a, 1996b) calls the upper-bounding implication an entailment, so (40a) above entails that B has at most three children (as well as entailing that he has at least three). Although I go along with Horn and the other neo-Griceans in seeing this as the result of a pragmatic inference (hence not an element of semantically encoded meaning), I find van Kuppevelt's use of the term "entailment" pertinent. It highlights the point that these inferences contribute to the proposition expressed by the utterance rather than functioning as implicatures. The concept of "entailment" has been used to characterise (a) the relation between the semantically encoded content of a sentence and other sentences that follow analytically from it, and (b) the relation between the proposition expressed by an utterance and those propositional forms which follow analytically from it. These two characterisations have often been used interchangeably, but once the gap between semantic encoding and the proposition expressed is recognised, a gap plugged by pragmatically derived material, it is plain that these are not equivalent characterisations. The set of (a)-type entailments of any sentence/utterance is a proper subset of the set of (b)-type entailments. Van Kuppevelt (1996b)'s talk of scalar implicatures as topic-dependent entailments involves use of the term "entailment" in the second sense only; that is, they are entailments of the proposition expressed by the utterance whose existence depends on a prior pragmatic inference. This may sound somewhat paradoxical - an entailment which arises only as a result of a pragmatic inference - but if the term "entailment" is to be used of those propositional forms whose truth follows from the truth of the proposition expressed, then there will be many entailments that are pragmatic (or context-dependent). Whether or not it is wise to use the term "entailment" in this way is a separate matter.

I conclude this section by reasserting the distinction between the role of the upper-bounding pragmatic inference in the interpretation of number terms and its role in the interpretation of the other quantity terms, the point made first by Sadock (1984) and then, in more detail, by Horn (1992). Although Scharten (1997) wants to give them a unified account in terms of her semantic process of "exhaustive interpretation", she gives many examples which indicate that the case for the two-sided interpretation of the non-number terms being truth-conditional is much weaker than the case for the number terms. For instance, she juxtaposes and discusses the following two examples, where in both instances the inference inducing term is in comment position (pp.66-67):

(42) Q: How many pupils are there in your class?
A: 31.

(43) Q: What is your profession?
A: I am an architect.

She says of example (42) "If it turns out that there are 33 pupils in the speaker's class, then he will not have spoken truthfully." But in her discussion of the non-numerical example (43), which she takes to be also a case of exhaustive interpretation, she is far less categorical (rightly so, in my view): "Here again, if it turns out that the answerer is also a practising doctor, then she may not have spoken the whole truth, depending on relevance criteria in the situation at hand ... The answer [ ] may well be called true but it is incomplete ...". The implication seems to be that the inference (if there is one) that A is just an architect (that is, she does not have any other occupation) does not affect the truth-conditions of the utterance; that is, it is an implicature.

4.2. *The linguistic semantics of number terms*

Having established that the bilateral understanding of cardinals is an element of truth-conditional meaning, we still have the thorny issue of the semantics encoded by the number terms, that is, what the input to the pragmatic inference is. There are at least the following possibilities: (a) a polysemy analysis, according to which all cardinal terms have three senses: "at least n", "at most n", "exactly n"; (b) a univocal analysis on which the cardinals encode "at least n"; (c) a univocal analysis on which the cardinals encode "at most n"; (d) a univocal analysis on which they encode "exactly n"; (e) a sense-general analysis according to which they do not encode any of the meanings they can be used to communicate, but rather they encode a sense which is weaker than any of them, and from which each can be pragmatically derived. Options (a) and (c) will be considered only briefly and then set aside. I'll argue against (b), which is the dominant view, that of the neo-Griceans, and then wrestle somewhat inconclusively with (d) and (e), which are the two strongest possibilities, in my opinion.

The ambiguity/polysemy view is not favoured by many (though Richardson & Richardson (1990) is an exception); in fact, the arguments against it are few and not very compelling (Occam's Razor is usually brandished), but I will go with the radical pragmatic flow for the time being and assume that if we can mount a nice pragmatic account of how various communicated meanings are derived from a single encoded meaning then that is preferable to postulating a range of senses.

Possibility (c), a univocal analysis on which the cardinals encode "at most n" has not, to my knowledge, been proposed by anyone. However, if an "at least n" pragmatic inference could be shown to occur - and this seems quite feasible since it is merely a reversal of the usual neo-Gricean semantics and upper-bounding implicature - then the, arguably, most common understanding, "exactly n", would be accounted for (as the combination of encoded "at most" and inferred "at least"). I suppose this has not been proposed because of the standard unified approach to all scalars in terms of entailment scales: just as *most* entails *some* and is compatible with *all*, so *four* entails *three* and is compatible with *five*, hence the "at least" or lower-bound semantics has seemed the correct unitary semantics for all cases. However, now that the special nature of cardinals among scalars has been generally accepted, there is much less impetus to opt for a unitary (lower-bound) semantics. So-called scale reversal in the case of the cardinals has been recognised since Horn (1972), and its failure to carry over to the inexact scalars was noted by Sadock (1984, 143):

(44) a. That golfer is capable of a round of 100 (and maybe even 90/*110).
b. She can counter most of the arguments (and maybe even *some/all).

An "at most n" semantics for number terms seems to me to be no worse (or better) than an "at least n" semantics. What could not be accounted for on such a semantics is how the "at least n" understanding is reached, since it would require that just the pragmatically inferred lower bound

was communicated, the semantically encoded meaning having been replaced or cancelled.

Carston (1985, 1990/95) presented the converse of this last point in arguing against the favoured neo-Gricean semantics, option (b): from an encoded lower bound ("at least n") it is not possible to derive the interpretation on which the number terms are understood as upper bounded ("at most n") as in (45). This third reading for number terms has been largely ignored by the neo-Griceans, who have been concerned only to show how a bilateral ("exactly n") understanding is derived.

(45) a. She can have 2000 calories without putting on weight.
b. The council houses are big enough for families with three kids.
c. You may attend six courses (and must attend three).

Given the pragmatic resources of the neo-Gricean system, there is no obvious way to effect the switch from the encoded lower bound to a communicated upper bound. It would seem to require a pragmatic inference to the upper bound which takes the lower bound semantics as its input and then cancels it. The only established cases where a Gricean analysis has this general profile, that is, an implicature with nothing actually meant (communicated) at the level of what is said, are rhetorical cases (metaphor and irony), which involve a flouting of the maxim of truthfulness. These bear very little resemblance to the uses of number terms in (45).

There are a range of further observations that militate against this "at least" semantics. Koenig (1991, 141-2) says of example (46) "if *three* names the half-line equal or above 3, we cannot make any sense of the expression *more than three*", while it is intuitively clear that this is perfectly interpretable and refers to the scale points above 3:

(46) More than three people came.

There are corresponding predictions of redundancy and oddity with explicit modification of number terms by the words *at least* and *exactly*, problems that do not in fact arise. These points apply also, *mutatis mutandis*, to an "at most n" semantics for number terms. Scharten (1997, 52) points out that the neo-Gricean account predicts that, to the question *how many children does John have?*, *three* and *exactly three* are the correct, appropriate answers, while *at least three* is not. "However, the facts are the other way around: *at least three* and *three* are appropriate, *exactly three* is not". The arguments could be proliferated, but I think enough has been said to establish the point: it is false that cardinal number terms have a lower bound semantics.

So the two options left in play are: (i) a semantics which is truth-conditionally equivalent to "exactly n"; (ii) an underspecified (general) sense which is weaker than, but compatible with, all three interpretations, each of which has to be pragmatically derived. Option (i) is favoured by Sadock (1984), Koenig (1991), and Scharten (1997); option (ii) is favoured by Carston (1985), Atlas (1990, 1992), and Verkuyl & van der Does (1995). (As far as I can tell, Horn (1992) and van Kuppevelt (1996a) do not commit themselves on this issue.)

Sadock (1984, 143) advocates an exact meaning for the exact quantifiers (the number terms) and says "the pragmatic principle involved in their interpretation as one-way unbounded is one of loose-speaking, the same principle that allows us to describe France as hexagonal ... A speaker using *three* to indicate 'three or more' would then be conveying less than his words imply, rather than more." But this analogy is not very convincing. A much more striking analogy is between the "hexagonal" example and cases where the number term is patently used as an approximation as in (47a), paraphraseable as (47b), parallel with (47c), rather than with examples like (47d) ("at least three") and (47e) ("at most ten"):

(47) a. She earns a thousand pounds a month.
b. She earns roughly a thousand pounds a month.
c. France is roughly hexagonal.
d. Mary needs three A's to get into Oxford.
e. You may take ten books home.

Koenig (1991) makes a similar semantic proposal: cardinals should be given a "punctual" semantics, but in his view the "interval" interpretations (such as "at least n") are, in many instances, arrived at by constructional semantics. An example of this is the distributive reading of a plural noun phrase which allows an "at least n" understanding (exemplified in (48b)) as opposed to a collective or group reading, which requires the "exactly n" interpretation (exemplified in (48a):

(48) a. Three boys together carried a sofa up the stairs (*in fact four did).
b. Three boys hurt themselves on the obstacle course, in fact four of them did.

From the fact that three boys together brought a sofa up the stairs one cannot deduce that two boys together carried a sofa up the stairs, nor is it compatible with the possibility that four boys were involved in the operation (these facts are captured by the punctual semantics for number terms). On the other hand, from the fact that three boys hurt themselves, one can deduce that two boys hurt themselves, and it is compatible with the possibility that more than three hurt themselves; this has the look of the examples that fall under the neo-Gricean account of scalar terms.

This effect of the collective/distributive distinction is also noted by Horn (1992, 174) and Atlas (1990, 1992): they take it to be a product of pragmatic inference, for Horn in the derivation of the collective "exactly" case and for Atlas in both cases (the semantics of number terms being, in his view, weaker and more abstract than either of these readings). Koenig (1991)'s view is that the two readings are entirely a matter for semantics; the interval interpretation in (48b) is reached at the sentential level by a compositional semantics operating on a lexical semantics for number terms, which is the ordinary mathematical value, and a distributive reading of the noun phrase, a process which he spells out explicitly and formally. The result is given informally by the semantic paraphrases in (49):

(49) a. There is a set B, B is a set of boys of cardinality 3, B carried a sofa upstairs.
b. There is a set B, B is a set of boys of cardinality 3, for each b in B, b hurt himself on the obstacle course.

Koenig then completes the picture by employing a quantity maxim (which seems to presuppose a volubility principle (1991,147)) to derive an upper-bounding implicature in the distributed case, since this is often the preferred reading despite compatibility with higher values; in the case of (48b) without the follow-up clause, this would be "No more than three boys hurt themselves".

Scharten (1997, chapter 3), building on the views of Seuren (1993), takes a position on the semantics of number terms which is quite similar to Koenig's. She advocates what she calls a weak bilateralist semantics, according to which the semantic value is given by the set of sets of the appropriate cardinality, so, for example, the lexical item *three* is the set of sets of cardinality 3:

(50) $[[three]] = \{ X : |X| = 3 \}$

The linguistic expressions *three* and *exactly three* are truth-conditionally equivalent (though, presumably, not necessarily pragmatically equivalent, a matter she does not discuss). This is the understanding that arises in comment position in the information structure of an utterance. The weaker "at least n" interpretation, which is common in topic position, involves an existential quantifier operating over the set cardinality semantics.

First, it has to be noted that she sets aside a range of other examples in which the number terms have a one-sided reading, either "at least" or "at most", (whether in topic or comment), but notes that these tend to arise when the number term is interacting with a modal predicate. This is evident in the "at most" examples in (45) above and the "at least" examples in (51):

(51) a. In Britain you have to be 18 to drive a car.
b. Mary needs three A's to get into Oxford.
c. Women with two pre-schoolers are eligible for the welfare benefit.

Intuitively, when the issue is one of what is permitted/allowed, what is relevant is an upper limit, and when the issue is one of what is required/necessary, what is relevant is a lower limit. In similar vein to Koenig, she says that the different interpretations of the number terms may be "forced by the semantics of other lexical items in the sentence". It doesn't seem implausible that the right semantics of the right modal operators operates (in some way yet to be specified) on the bilateral semantics of the cardinals to weaken them in the one direction or the other, though she does not attempt to show this.

Her main thesis, which is pitched against the neo-Gricean implicature account, is that two-sided readings of scalars are to be explained by a **semantic** process of exhaustive interpretation which is triggered by the position of the scalar term in the information structure of the utterance; as already mentioned, in comment position scalars are exhaustively interpreted (given an "exactly" reading), in topic position they are not. However, the postulated semantics of number terms is equivalent to the "exactly" understanding anyway, so, in fact, no process of exhaustive interpretation is required anywhere; rather, what is required is some mechanism by which scalars in topic position, as in (52 (A1)) are weakened to a merely lower bounded interpretation:

(52) Q1: Who owns four sheep?
A1: **John** owns four sheep.
<Q2: How many sheep does John own?>
A2: (in fact) he owns **twenty**.

(comment in bold)

The idea seems to be that recognition that the scalar term is in topic position, either due to a previous topic-forming question (as in Q1 here) or to intonational clues, triggers a process of embedding the set cardinality semantics within the scope of an existential quantifier. I omit the technical details of how this works, but it is not too difficult to see how this leads to the "at least" understanding, since the existential is defined in the usual way so that it means "there is a [= at least one] ...".

Finally, Scharten has to account for the fact, which she acknowledges, that the preferred interpretation of numerals not in comment position is still "exactly n"; that is, A1 in (52) would most often get an "exactly four" interpretation in the absence of a follow-up such as A2. Both Koenig (1991) and Fretheim (1992) recognise this too and, although they favour a semantic treatment for the "exactly n" interpretation in those contexts where it affects truth-conditions, they turn to pragmatics for the preferred upper-bound interpretation in those cases where they take the

truth-conditional content to be "at least n", Koenig using a quantity maxim and Fretheim using relevance theory. However, following Seuren (1985, 1993), Scharten eschews talk of pragmatics altogether and adverts to a further discourse semantic principle which is "responsible for turning the truth-condition 'at least one entity satisfying the conditions specified' into the stronger condition 'exactly one entity (etc.)'" (Scharten 1997, 79); this is a principle of default interpretation, according to which "specific reference is preferred over non-specific reference".[14]

As it stands then, the account has three ways of accounting for the two-sided interpretation: by the intrinsic semantics of cardinal number terms, by the semantic process of exhaustive interpretation and by a further principle of default interpretation which, in effect, acts to undo the work done by a semantic process of embedding the bilateral lexical semantics in a higher-level existential quantifier. Setting aside this final baroque twist, it seems that, at the least, either the concept of exhaustive interpretation is redundant or a different lexical semantics for number terms is called for, one which gives this process something to do.

Carston (1985), endorsed by Atlas (1990, 1992)[15], argues for a semantics of number terms which is neutral among the three interpretations, "at least n", "at most n", "exactly n", so that they don't have any one interpretation until they are placed in a particular sentential context, and sometimes a wider context is necessary. In other words, number terms are semantically incomplete; the semantics of *three*, for example, can be conceived of as follows:

(53) [X [THREE]]

This representation overtly requires that material be supplied pragmatically to instantiate the variable X; that is, the necessity of a process of pragmatic enrichment is signalled in the logical form (semantic representation) of the utterance. Such an analysis has been advocated for other constructions, including possible genitive relations and quantifier domains, (by, for instance, Recanati (1989)):

(54) a. John's book is excellent.
a'. [The book [in relation X to John]] is excellent.
b. Everyone went to Paris.
b'. [Everyone [in domain X]] went to Paris.

Here, a process of pragmatic instantiation of the variable X is obligatory; the context dependence of the interpretation is linguistically indicated. We don't have a complete proposition (a determinate set of truth conditions) until that slot is filled, or "saturated" as Recanati puts it).

Many of the arguments used to support this view have been surveyed above, though it has to be admitted that while they weigh strongly against the standard neo-Gricean "at least n" semantics, most of them do not decide between a sense-general treatment and the two-sided or set cardinality position. Consider, for instance:

(55) Q: Does she have three children?
A1: No, she has two.
A2: ? Yes, (in fact) she has four.
A3: No, she has four.

According to the standard Hornian/Levinsonian account, (A1) and (A2) should be entirely natural, unmarked responses, while (A3) should be somewhat marked, involving a metalinguistic negation (targeting the scalar implicature "at most three"). But this is not how the judgements go. Both the

sense-neutral and the exact (two-sided) semantics for number terms are compatible with the acceptability of the two negative responses; on the sense-neutral account, the "X" in [X [THREE]] is instantiated as "exactly", which is the most natural enrichment in this case. So on both accounts, the negative responses can be based on either the knowledge that she has only two children or the knowledge that she has four children, and the negation in both of the responses is straightforwardly descriptive. Regarding an explanation for the somewhat marked response in (A2), the sense-general account seems to have a slight edge over the exact semantics. What a hearer of (A2) takes its speaker to be agreeing with, in the first instance, is that the woman referred to has exactly three children, but this is followed up by *she has four*, giving a contradiction. On the sense-general account, since the "exactly" was a pragmatic enrichment, it is possible to go back, as it were, and reprocess, instantiating the variable X in a different way, as "at least", thereby making the utterance as a whole consistent. There is extra processing here and the interpretation is not smooth, which is consonant with intuitions. Starting with an "exactly" semantics (of the Sadock/Koenig/Scharten sort) accounts for the processing blip too, but it is much less clear how the process of repair or reanalysis is to be explained. Sadock (1984) would see it as a process of pragmatic loosening but, as argued above, this bears little resemblance to the *France is hexagonal* example he cites by way of analogy, or to any other cases of loosening in the literature, so although this account might prove right ultimately, there is a chunk of analysis yet to be devised.

There is one other consideration which provides some support for the semantic generality thesis for number terms. It has already been mentioned that number terms can be explicitly modified by *at least* and *at most* without giving rise to any redundancy or oddity, as in (56a) and (56b); this points away both from the "at least n" semantics and from a polysemy account, but it is compatible with the two possibilities still on the table.

(56) a. If you get *at least* three A grades, you'll be admitted to Oxford.
b. He can consume *at most* 2000 calories a day without putting on weight.
c. There are *exactly* nine students taking the semantics course.
d. I took *exactly* six cigarettes with me, gave *exactly* one to Ed and *exactly* two to Ed, so I still have *exactly* three.
e. Oxford is *approximately* sixty miles from London.

Explicit modification by *exactly* as in (56c) also seems fine, which favours the sense-general account, and while (56d) may seem a little strained due to the pedantic repetition of *exactly*, it does not have the smack of semantic redundancy that makes the following cases not merely strained but downright anomalous:

(57) a. John is an unmarried bachelor.
b. Mary is an adult spinster.
c. The dead murdered woman was found in a ditch.

The problem with (56d), if there is one, is that it spells out repeatedly something which is very readily pragmatically inferable. An addressee would inevitably look for some effects, derivable from this insistent encoding of *exactly*, which would not be communicated by the unmarked use of the bare number terms. An appropriate context might be one in which, because cigarettes are in short supply and great demand, there is special significance attached to how many cigarettes an individual has at any given moment and so to giving any of them away; in the absence of any such accessible context, the hearer will be left with the sense either that the speaker is trying to make

some point that is escaping him or that she is a painful pedant. It does not, I think, leave one with the sense that the speaker has been semantically redundant or doesn't understand the full meaning of some of his words, as would be predicted by an "exactly" semantics for number terms. Finally, the variable slot in the neutral semantics might also be filled by an *approximately* modification, as in (56e). This is rather slender evidence, however, and it seems to me that both the sense-general and the two-sided punctual semantics remain live options, though each has its problems and neither is fully worked out.

4.3. *Summary and residual points*

I conclude section 4 with a brief summary. Among those sets of elements that support the sort of inference that has been classically assumed to issue in scalar implicature, the number terms seem to be special: (a) they can be understood as punctual ("exactly n") or as communicating an interval ("n or more", "n or less"), so (b) their lexically encoded meaning seems not to be that of a lower-bounded interval, and (c) whichever of these three possibilities is communicated on a particular occasion of use, it is a part of the proposition expressed, that is, truth-conditional. The demise of the neo-Gricean semantics removes the motivation for seeing the inferential process here as based on the Horn/Levinson Q-principle. Rather, the saturation or enrichment process can be seen as relevance-driven, the choice of enrichment on a particular occasion of use being determined by the drive to find an interpretation which yields an adequate range of effects for no unnecessary processing effort.

The non-numerical scalar terms, on the other hand, (a) can be understood either as bilateral (e.g. "some but not most/all") or as lower-bounded (e.g. "some and perhaps most/all"), but not as upper-bounded (there is no scale reversal), so (b) some of them may well have a lower-bound semantics as assumed by the neo-Griceans, and (c) their strengthened interpretation (by an upper-bounding inference) does not seem to be truth-conditional but to be a case of conversational implicature. How relevance theory might account for these is considered in the next section.

There are still other cases of linguistic expressions, perhaps the bulk in fact, where nothing can be concluded about their semantics from the fact that they may give rise to scalar effects (Koenig (1991, 151) argues this point forcefully). These are cases where their scalarity is wholly a matter of pragmatics (a possibility stressed by Fauconnier (1975)). Examples involving proper names are perhaps the clearest, since there is no temptation to attribute some sort of scalar boundary to their semantics:

(58) A: Hussein is a Mussolini, if not a Hitler.
B: He's a Mussolini.
implicating: He is not Hitlerian.

However, any set of objects, properties or events may be viewed hierarchically in a particular context and so give rise to scalar effects. The following attested example from Horn (1989, 241) demonstrates this (and so do many of the examples in Hirschberg (1985/91):

(59) Overt antifeminism, if not homosexuality, may be the result of such experience in the male.
(from *The Parenting Advisor*, on the failure to shift gender identification from mother to father.)

Homosexuality does not entail *antifeminism* by virtue of linguistic meaning, but the *if not* clause used here induces a scale on which the two are ranked, with *homosexuality* higher on the scale

than *antifeminism.* In such a context, predicating antifeminism of someone might implicate that he is not (however) homosexual.

In the next (and final) section, I return to the issue of the pragmatic principle(s) responsible for the inferences that give an enriched proposition expressed or a conversational implicature. I focus now on the account offered by Relevance Theory, according to which all the work is done by a single Communicative Principle of Relevance, and compare its predictions with those of some of the various relevance-constrained informativeness maxims discussed earlier.

5. Relevance theory and scalar inferences

Many authors have noted that Relevance Theory has had little to say about the issue of scalar implicature; some have gone further and asserted that the theory is intrinsically incapable of accounting for this sort of inference (Levinson (1989, 466), Welker (1994, 80)). If this were true, it would indeed by a shortcoming of the theory since this has been one of the central topics of post-Gricean inferential pragmatics. It has, in fact, been addressed by Sperber & Wilson (1987, 748), Carston (1990/95) and, most recently, by Sperber & Wilson (1995), who claim to be able to account quite smoothly for scalar implicature. In the next two subsections, I shall review their discussion, extend it to some further examples and compare it with some of the approaches surveyed above. In the final subsection, I shall look at a scalar implicature example which has been presented by Matsumoto (1995) as a counterexample to the predictions of relevance theory; I think he is wrong.

5.1. *Relevance and the communicator's abilities and preferences*

Recently within relevance theory, there has been a small revision in the formulation of the key concept, "optimal relevance", a revision which naturally affects both the Communicative Principle of Relevance, which says that every utterance (more generally, every act of ostensive communication) communicates a presumption of its own optimal relevance, and the comprehension strategy that hearers are justified in following on the basis of this presumption. In most respects, this is a case of the formal definition catching up with actual practice in applying the theory, but it is useful to look at the two formulations to see how the literal application of the new one extends the predictive power of the theory beyond that of the literal application of the earlier one. One area of pragmatics to which this extension applies is scalar implicature.

Presumption of optimal relevance (original)

(a) The set of assumptions **I** which the communicator intends to make manifest to the addressee is relevant enough to make it worth the addressee's while to process the ostensive stimulus;
(b) The ostensive stimulus is the most relevant one the communicator could have used to communicate **I**.

(Sperber & Wilson 1986/95, 158)

Presumption of optimal relevance (revised)

(a) The ostensive stimulus is relevant enough for it be worth the addressee's effort to process it.
(b) The ostensive stimulus is the most relevant one compatible with the communicator's

abilities and preferences.

(Sperber & Wilson 1995, 270)

I'll try to highlight the differences between the two versions, before going on to see how the revised one compares with versions of the first Quantity maxim in predicting a range of cases of implicature including, in particular, scalar implicatures. Let's start with clause (b) where the change is particularly clear, in the relativization to the speaker's abilities and preferences. There was always something a little dubious about clause (b) in the original formulation: can communicators really be credited with finding the best possible (the least effort-demanding) utterance for achieving their communicative ends? Surely not, and, in fact, in most actual relevance-theoretic analyses in the intervening years, this ideal has been somewhat hedged by talk of a stimulus which required no "unjustifiable" or "gratuitous" effort from the hearer (in achieving the expected level of effects). Clearly, vocabulary limitations, linguistic idiosyncrasies, preferences for certain turns of phrase, politeness concerns, political correctness constraints and on the spot performance problems often lead a communicator to produce an utterance stimulus which falls short of perfect economy from the hearer's point of view, but this more-than-minimal effort required is not gratuitous, not unjustifiable: it is the result of the speaker's (flawed) abilities and (possibly conflicting) preferences. So this was always recognised within the framework, but it is made explicit in the revised presumption.

Now looking at both clauses together: in the original version, clause (a) says the level of contextual effect is (at least) sufficient to warrant the hearer's attention, and clause (b) says that the stimulus chosen by the speaker is the most economical (the least effort-demanding), that she could use to achieve those effects. There are two closely related points to note here: (i) there seems to be an asymmetry between effect and effort (sufficiency of the one, minimality of the other), which is not obviously correct, and (ii) while sufficiency of effects, captured in clause (a), is the lower limit of what an utterance should provide in order to justify its imposition on the hearer, able and cooperative speakers often go well beyond mere sufficiency, a fact that is not reflected in the presumption. In the revised version, effort and effect are treated symmetrically: clause (a) captures both sufficiency of cognitive gain and sufficiency of cognitive economy (that is, the stimulus is low enough in the effort expenditure it demands from the hearer so as not to detract from the level of overall relevance); clause (b) sets an upper limit on both, an upper limit that captures the fact that an utterance may be more than merely adequate in its contextual effects, though obviously it cannot go beyond what a communicator is able and willing to offer, and, as already discussed, an upper limit on how undemanding of hearer effort an utterance stimulus can be. The effort-effect symmetry in this formulation is signalled by the shift from talk of "the set of assumptions I" to consistent use of "the ostensive stimulus" throughout.

The comprehension strategy warranted by the presumption of optimal relevance is as follows:

(i) consider possible interpretations in their order of accessibility (i.e., following a path of least effort); and
(ii) stop when the expected level of relevance is achieved (or appears unachievable).

Where the two different definitions of optimal relevance could make a difference in the implementation of this strategy is in what constitutes the "expected level of relevance". According to the first version, the expected level is one of "sufficiency", enough effects to justify the request for the hearer's attention. What this amounts to is that the effects afforded by the utterance should

be greater than those of the other phenomena in the environment that the hearer could have been attending to, so what constitutes sufficiency will vary considerably from one set of circumstances to another. According to the second version, the hearer is entitled to expect a level of relevance which is at least sufficient and which is as high as is compatible with the speaker's means and goals. For the most part, this shift makes no difference to predictions regarding hearers' interpretations. Consider the following example, in which Ann, Bob and John are siblings; Ann rings Bob and says:

(60) John has won the jackpot.

Now let us suppose, as is highly likely, that the hearer, Bob, knows several people called John, and that some of these individuals are mutual acquaintances of Bob and his sister. Still, the first interpretation that comes to Bob's mind (i.e. the most accessible one) is that his brother John has won the jackpot; this has a large range of cognitive effects, well over the sufficiency threshold. Learning that an old friend of his and Ann's, John Allan, had won the jackpot would have had fewer cognitive effects, but still quite enough; it too would have been sufficiently relevant, though not as relevant as the interpretation he derives. This sort of case is captured equally well by the two slightly different criteria that follow from the two different formulations of optimal relevance. It is obvious that the first interpretation accessed by Bob meets the expected level of relevance that comes from the revised presumption. But it also meets the "at least sufficient" level of the unrevised version; it surpasses it considerably and there are other possible interpretations that are closer to the lower limit, but that does not affect the comprehension process, which involves testing interpretative hypotheses in their order of accessibility. It is this effect-packed interpretation which is both the most accessible and which meets the criterion.

So what sort of case could come out differently on the two versions? It would have to have the following structure: the first interpretation accessed by the hearer would be sufficiently relevant but would not be the most relevant one compatible with the speaker's means and goals. On the unrevised presumption, such an interpretation would be predicted as the chosen one. The revised presumption would predict that it is rejected in favour of another more relevant interpretation, that is, one with a further range of effects for the hearer. This will not constitute a wholesale rejection, but rather an augmentation ("there's more to be had here"), such that the effects of the merely sufficient interpretation constitute a subset of the effects of the interpretation which is the most relevant one compatible with the speaker's means and goals. Assigning reference to different individuals and choosing different senses of ambiguous expressions generally give rise to interpretations which are not in this subset relation with each other; so cases where the two different presumptions make different predictions about these pragmatic processes are likely to be uncommon (however, see Breheny (1997) for some interesting possibilities). The most likely locus of difference is at the level of implicature. We will see an example with this sort of shape in section 5.2, where the revised presumption correctly predicts a scalar implicature which is not predicted by the earlier version.

Before that, let's consider some examples where the revision makes for a more streamlined account of the derivation of implicatures. Sperber & Wilson (1995, 273-4) discuss the well-known Gricean example in which A and B are making plans for a trip to France, and A would like to visit their old acquaintance Pierre. The following exchange takes place:

(61) A: Where does Pierre live?
B: Somewhere in the South of France.
implicating: B does not know where in the South of France Pierre lives.

Given the original presumption of optimal relevance, the implicature here would be explained by there being a contextual assumption that B is willing to co-operate in helping A in his plans. This, together with the fact that B's reply is not relevant enough to answer A's question, gives rise to the implicature that she does not know exactly where Pierre lives. Sperber & Wilson go on to consider a different possible implicature of B's utterance:

(62) B is reluctant to disclose Pierre's exact whereabouts.

This would require a contextual assumption that B knows full well where Pierre lives, which, together with the fact that B's reply is not relevant enough to answer A's question, could give rise to this implicature. It may be that a further manifest assumption is required in each case to make the inferences go through, roughly: "if someone wants to help but doesn't, that's because they can't" for (61), and "if someone is able to help but doesn't, that's because they don't want to" for (62).

There is nothing wrong with these accounts, but the revised presumption gives a more transparent analysis, one on which the two implicatures flow more directly from the presumption itself, without the need to posit contextual assumptions concerning the relation between people's behaviour and their abilities and goals. The "unable to say" implication follows directly from the proposition expressed by B, the contextual assumption that B would like to be more specific about where Pierre lives and clause (b) of the revised presumption (she wants to help, so she is implying that she cannot). Similarly, the "unwilling to say" implication follows directly from the proposition expressed by B, the contextual assumption that B is able to be more specific and clause (b) of the revised presumption (she is able, so she is implying that she prefers not to). In either case, if it is mutually manifest that this implication increases the relevance of her utterance, it is an implicature (that is, the hearer is justified in taking it as part of the intended interpretation).

Let's now compare this with the Gricean account. The implicature in (61) is discussed by Grice (1975) as a case involving a clash between his first maxim of quantity (requiring sufficient informativeness) and his second truthfulness maxim (requiring an evidential base for what is said). In obeying the second of these the speaker cannot help but infringe the first, and this gives rise to the implicature: the speaker is unable (through lack of knowledge) to give the required information. That is, the implicature comes directly out of the interaction of the maxims; this is, arguably, a neater account than that afforded by the unrevised presumption of optimal relevance. On the revised version, however, we get an equally direct and streamlined account, though one which is quite different from Grice's.[16] With regard to the implicature in (62), the relevance-theoretic account is clearly better, since this sort of implicature doesn't seem to be derivable at all using Grice's system. The problem is that it involves the hearer in recognising the absence of speaker cooperation and in his scheme, whatever maxims may be violated, the ultimate interpretation of an utterance must be such that the assumption of speaker compliance with the overarching Cooperative Principle (CP) is preserved. It can only be a case of what Grice calls "opting out", which does not give rise to implicatures at all. I assume the same problem arises for Welker (1994), whose generative account of implicature runs on a formalised revision of the Cooperative Principle, which assumes there is a mutual conversational goal and the speaker's aim is to bring the common ground closer to that goal (see footnote 8).

These examples bring out some advantages of the relevance-theoretic approach over the Gricean. "Abilities" enter indirectly into the Gricean system through the maxims of truthfulness and informativeness (speakers sometimes simply cannot truthfully provide the information required), but "preferences" do not feature, the system assuming a level of willingness in speakers that is simply not always present and not always expected by hearers to be present. The neo-Griceans (Levinson, Horn, Hirschberg, Matsumoto) seldom mention the Cooperativeness Principle

when they give implicature derivations based on their own systems of maxims. If it is tacitly assumed to be in force, then the problem for Grice just outlined arises for them too; if it has been abandoned, there is still no obvious way of deriving the "unwilling to say" implicature, since none of their maxims provides a place for the personal interests and goals of the speaker. So there is no possibility of a clash between required informativeness and speaker willingness, parallel to the clash between required informativeness and evidencedness, which underlies the "unable to say" implicature.

It is worth noting that the "unwilling to say" sort of implicature, although hardly discussed in the literature, is not a marginal phenomenon; here are a few examples, based on attested cases I have heard since thinking about this matter:

(63) A: When will you be back?
B: When I'm ready.

(64) A: Which of your colleagues support the strike?
B: Some of them do.

(65) A: How many clients do you have?
B: A number.

These are all cases where unwillingness to give some information could reasonably be assumed to be implicated. A has asked for quite a specific level of information and has received something much less specific from B, who, in each case, patently has the more specific information at her disposal; these are highly uncooperative responses which seem to be geared to warning A off pursuing the matter any further. It's a major gap in the Gricean and various neo-Gricean accounts that they have no way of predicting this sort of implicature.

To end this subsection, let us consider a very different sort of example, taken from Welker (1994, 51):

(66) A: I'm having a dinner party and I need four more chairs.
B: John has six chairs.

She discusses this in the context of a survey of Grice's conversational maxims, so B's response is seen as a violation of the second maxim of Quantity ("Do not make your contribution more informative than is required"). More than the required information, that John has four chairs, has been given, which forces the hearer to look for an implicature so as to preserve his assumption that the Cooperative Principle is being observed. In her view, "the additional information is used to reinforce the likelihood that John will loan the chairs -in other words, to provide evidence that the plan B is suggesting is a good one for A to adopt. This relies on an assumption on B's part that John's having *more* than four chairs would make John more likely to loan A the four chairs. So the implicature arises here that John is especially likely to loan the chairs."

My interest in this example is that I think it is particularly nicely accounted for by the revised presumption of optimal relevance which makes it explicit that sufficiency of effects is merely the lower limit and a hearer is licenced to look for a higher level, modulo the speaker's goals and means. B knows that John (who, let us suppose, is also coming to the party) has six chairs and that A is in need of four chairs; she can utter truthfully either (i) *John has four chairs*, or (ii) *John has six chairs*, neither requiring more effort from her than the other. If she had chosen to utter (i) it would certainly have been sufficiently relevant to A. Why, then, did she choose (ii)?

Because there's a good chance that it will be more relevant to A than (i): it will have further effects, perhaps along the lines suggested by Welker, with a negligible, if any, increase in processing effort for the hearer. So B's choice of this utterance is entirely in line with the revised presumption of optimal relevance.

It is unclear how the Gricean analysis of (66), involving the second maxim of Quantity, is supposed to go; it doesn't seem to be a case of a clash with another maxim, nor does it have the rhetorical flavour of a maxim flouting, and the sort of implicature that Welker suggests it gives rise to does not reinstate its preservation at another level. In any case, violation of this maxim seems frequent and ordinary, which suggests that the maxim itself is simply wrong:

(67) A: Excuse me, where is Professor Smith's room?
B: It's on the next floor up, but he isn't in today.

(68) A: I need a pen.
B: There's pens, paper and paper-clips in the left hand drawer of the desk.

Human cognisers often anticipate what an interlocutor may find relevant, without having to have it explicitly signalled to them by a previous question, and they choose to be generous in their offerings. This is nicely captured by the second clause of the revised presumption of optimal relevance which allows that able and well-disposed speakers may, on occasion, give information which goes beyond that which is strictly required.

5.2. *Scalar implicature: optimal relevance versus quantity principles*

In this section, I'll concentrate on a variety of cases involving the paradigm scalar implicating term, *some*, and compare the predictions made by the two definitions of optimal relevance and the two versions of the first maxim of quantity, Grice (1967)'s sufficiency version, and the volubility version, espoused by most of the neo-Griceans (Harnish, Horn, Levinson, Hirschberg, Matsumoto). Sperber & Wilson (1995, 276-78) discuss the following two examples:

(69) A: If you or some of your neighbours have pets, you shouldn't use this pesticide in your garden.
B: Thanks. We don't have pets, but some of our neighbours certainly do.

(70) A: Do all, or at least some, of your neighbours have pets?
B: Some of them do.

Intuitions generally agree that (69) is a case where there is no scalar implicature along the lines that "not all of B's neighbours have pets" while (70) is a case where this implicature clearly arises. According to Sperber & Wilson (1995, 277), the unrevised presumption of optimal relevance, applied mechanically, predicts the nonoccurrence of the implicature in (69) but does not predict its occurrence in (70). The reason is that in both cases the proposition expressed by the utterance, that *at least some* of B's neighbours have pets, is relevant enough (has a sufficient level of effects).[17] The difference between examples (69) and (70) is that A's question in (70) makes it plain that while the information that at least some of the neighbours have pets is sufficiently relevant to him, it would be more relevant to him to know whether or not all of them do. The revised formulation explicitly recognises that an able and/or cooperative speaker (one whose knowledge states and personal concerns do not conflict with giving the hearer all the information

he would like) will, in fact, often provide more than the lower limit of sufficient effects, and that a hearer is thereby licensed to recover an interpretation that achieves this, provided the processing effort involved does not detract from overall relevance.

So the relevance-theoretic comprehension strategy could give two different results here, depending on whether the expected level of relevance in the criterion is as the original presumption of relevance says or as the revised one says. If the first interpretation accessed by the hearer is the "some but not all" interpretation then, of course, both versions accept this as the right interpretation, since it is over the level of sufficient relevance; this possibility seems quite likely to me, especially if B gives the word *some* heavy stress. However, the possibility that Sperber & Wilson are considering is that the first accessed interpretation is the weaker "at least some" understanding; while this is acceptable on the unrevised version, it will not be accepted on the revised version, since there is a more relevant interpretation compatible with the speaker's means and goals, one which incorporates the sufficiently relevant interpretation and goes beyond it. B's answer makes it manifest that she is either unable or unwilling to inform A that all her neighbours have pets. Either of these is possible and could lead to an implicature of either of the two types. Let us suppose, however, that it is mutually manifest that she intends to be entirely cooperative; then it follows from clause (b) that she is unable to give the more informative answer. There are two possible reasons for this, two ways in which she may lack the desired information: either she doesn't know whether all her neighbours have pets or she knows that not all of them do. Either one of these may be implicated, depending on mutually manifest assumptions about B's knowledge of the situation or, conceivably, there may be some indeterminacy regarding which is intended. If it is the weaker case, then this example falls together with the "South of France" example, where the implicature concerns the speaker's ignorance of the facts of the matter;[18] in the stronger case, the speaker is fully knowledgeable but the information she has is at odds with the information sought by the hearer.

Naturally, the accounts of Horn and Levinson correctly predict the scalar implicature in (70), since for them the scale <all, some> is automatically activated by the use of *some* and the default inference to "not all" goes through unless something explicitly blocks it. It is cases of nonoccurrence of the implicature, such as in (69), which present a big problem for them. This point is made very emphatically by Green (1995, 96-97), who discusses a range of examples, including the following:

(71) B: Are some of your friends Buddhist?
A: Yes, some of them are.

As Green says, there is no reason to suppose here that A is (scalar) implicating that not all of her friends are Buddhist. A gives B that information which B's question has made manifest would be relevant to him. Green elaborates the case further:

> ... suppose that A knows that B, in spite of only asking whether some of A's friends are Buddhist, would also be interested to know if in fact all, or most, of A's friends are Buddhist [since, let us suppose, B is gathering data about interactions amongst people of different religious groups]. And suppose that B is aware of the fact that A knows this. In this situation a stronger assertion such as '*Yes, in fact all of them are*', would be relevant. But it does not follow that in giving only the weaker answer, *'Yes, some of them are Buddhist'*, A is implicating that she is not in an epistemic position to make a stronger claim. It *would* be more generous for A to make the stronger remark - that all of her friends are. Yet A might have some reason for diffidence concerning this stronger point,

such as fear of being considered a Buddhist-groupie. Such a reason might prevent A from being as generous with her information as she might ...

Green's point is that the neo-Gricean Q-principle (= volubility) wrongly predicts a scalar implicature here, since there is a stronger relevant statement that the speaker could have made. Grice (1967)'s first maxim of Quantity, however, gives the right prediction; it says that a speaker should be at least as informative as is necessary/required (and *may* be more so). A's response is as informative as required by B's question, so there is no prediction that a scalar implicature (*Not all my friends ...*) is communicated.

Let us see how the Communicative Principle of Relevance, based on the revised presumption of optimal relevance, fares with this sort of example.[19] First, it is clear that A's response meets the requirement of the first clause of the presumption: A's utterance certainly does have sufficient effects for it to be worth the addressee's effort to process it, since it gives exactly the information B has asked for. Now, in the context that Green sketches, it is evident that there is a more relevant response that A could have given, concerning whether all or most of her friends are Buddhist; this would have more contextual effects for the hearer (B) and would cost him negligible further processing effort. Since A has chosen not to utter this, doesn't it follow that she must be communicating that *only some* (that is, not all or most) of her friends are Buddhist? This would follow from a presumption of *maximal* relevance (just as it would follow from a maxim of maximal informativeness, i.e. volubility). However, Green's context makes it plain that while the speaker has the *ability* to make the stronger statement, she *prefers* not to (she is afraid of being considered a Buddhist-groupie) and the hearer is aware of this. Hence the relevance principle correctly predicts that the speaker is not implicating that not all of her friends are Buddhist and that the hearer recovers no such assumption as part of what is communicated.

Given the details of the example, it is also not predicted here that the speaker is implicating (i.e. ostensively implying) an unwillingness to disclose the extent of her friendship with Buddhists. The hearer might derive this as an implication (a contextual effect) of her utterance, but unless it is mutually manifest that the speaker intends to make her unwillingness manifest, it cannot be taken to be part of what she has communicated. Had B's question been *Are all your friends Buddhist?* or *Do you have a lot of Buddhist friends?*, where it is made plain that a sufficiently relevant answer requires a stronger response than *Some are*, then in the same sort of context of assumptions about A's potential embarrassment, this answer might be taken to implicate that she is unwilling to say more and, even further, that B should not pursue this line of questioning further. Green's account does not address the "unwilling to disclose" type of implicature and he maintains the fundamental Gricean conversational principle of Cooperativeness.

However, his discussion of different speaker motivations on different occasions meshes well with the insights underlying the revised presumption of optimal relevance. In a general consideration of the rational bases of communication, he points out that it is sometimes in the speaker's interest that she make her hearer as well informed as possible (this may be useful to the speaker at some later stage), but, equally, in other instances, it runs against the speaker's interests that the hearer be maximally well informed (this may be used against her at some later stage) (Green 1995, 101). In other words, while volubility is sometimes judged advantageous to the speaker (hence, other things being equal, preferred) it is sometimes not (hence, other things being equal, dispreferred). There is no case, then, for promoting volubility to the status of a pragmatic maxim/principle.

Green reinstates the Gricean first maxim of Quantity, enjoining adequacy of information, while granting that a speaker may sometimes choose to go beyond the minimal required level of informativeness. But how could a hearer employing the Gricean system of maxims determine, on

any occasion of use, whether a speaker is being merely sufficiently informative or is moving closer to maximal informativeness? For Green, at least part of the answer lies in an appreciation of the kind of conversation or discourse that is taking place, since different kinds tend to require different levels of informativeness; he discusses in particular three sorts: informational exchanges, inquiries and debates. In other words, it is necessary to build in contextual assumptions about the aims of the discourse one is taking part in (for instance, in a debate a speaker can be expected to be unwilling to give up more ground than he absolutely must), and the functioning of the quantity maxim is taken to be relative to these. As far as I can see, there are at least the following three reasons for preferring the account offered by Sperber & Wilson's revised presumption of optimal relevance: (a) it can account for the "unwilling to disclose" sort of implicature, while the Gricean account cannot; (b) more generally, building into the principle considerations of speakers' abilities and preferences makes it sensitive, not only to type-of-discourse considerations, but also to more individual and/or transitory preferences (such as the speaker's diffidence in example (71) above); (c) while Green's account maintains the full set of Gricean maxims, including the unexplicated maxim of relation (relevance), the result of Sperber & Wilson's development of a theory of cognitive relevance is a communicative principle that is powerful enough on its own to account for the full range of Gricean implicatures, making any other principles, including a distinct quantity principle, unnecessary.

The well-known examples, discussed by Keenan (1976), of a society interacting under quite stringent constraints on informativeness are also, arguably, better accommodated by the principle of relevance than by any version of Grice's quantity maxim. In the Malagasy community she studied, it is the norm to reveal less information than the addressee wants. There are two reasons for this: (a) new information is a highly prized commodity and gives one such power that it is not to be readily given away, and (b) being the one responsible for communicating information, especially if that information concerns others, carries strong social sanctions, not only for oneself but for one's whole family. Here is a simple example of the sort of communicative exchange that this state of affairs gives rise to:

(72) A: Where is your mother?
B: She is either in the house or at the market.

B's reply here does not implicate that she, B, does not know whether or not her mother is in the house and does not know whether or not her mother is at the market, as it is standardly said to do in British or North American culture.[20] Keenan saw the communicative practices of this community as providing counterexamples to Grice's first maxim of Quantity, which is therefore, at best, culture-relative. If this were a problem for the maxim of sufficient informativeness, it would be all the more so for the neo-Griceans' volubility principle. In fact, it is not really a worry for the Gricean view. An example like (72) is a case of "opting out" of observation of a maxim, a possibility that Grice explicitly allowed for and which is, plainly, a rational thing for a member of this society to do in a great many instances. Opting out is rational behaviour in certain circumstances in European societies too, when giving a piece of sought information might have unpleasant consequences for oneself or for others.

Nonetheless, the nonoccurrence of these quantity implicatures does follow more directly and smoothly from the Communicative Principle of Relevance; there is no need to postulate a special category of opting out (in fact, one cannot opt out of this sort of cognitively based principle). It follows directly from considerations of a speaker's preferences (in this case, a reflection of community-wide preferences) that the speaker is not communicating an inability to be more specific. I don't think that, in this case, B implicates a reluctance to be more precise

either, since avoidance of specificity in an exchange like this is a general background assumption in this culture. Perhaps, if A is the alien anthropologist still new to the ways of the Malagasy, and if she is quite certain that B knows where her mother is, she will take B to have implicated that she is unwilling to be more specific; then there will be a cross-cultural misunderstanding.

While we're on the disjunction case, recall the treasure hunt example (Grice 1978, 116-7) which, in any culture, does not communicate ignorance of the truth value of the disjuncts:

(73) The prize is either in the garden or the attic.

Here the speaker's preference, dictated by the game that everyone concerned is sharing in, is to withhold knowledge she has and is known by the participants to have. She does not implicate the lack of more specific knowledge, but nor does she implicate an unwillingness to disclose, since knowing but not disclosing is a situational given. I assume that it is obvious by now how the revised presumption of optimal relevance accounts for the absence of these implicatures, despite the fact that there is a more informative utterance the speaker might have produced.

In these last two subsections, I have gathered together a range of cases: some where the presence of a scalar term such as *some* gives rise to a scalar implicature, some where it does not, some where an "unable to say" implicature arises and some where an "unwilling to say" implicature arises. Neither Grice (1967)'s first quantity maxim, operating within his overarching Co-operative Principle, nor the neo-Griceans' volubility version of informativeness, covers the full gamut of cases. The Communicative Principle of Relevance, based on the 1986 definition of optimal relevance, does much better, but in a few instances, such as (70) above, it requires somewhat gratuitous seeming extra contextual assumptions. The advantage of the revised definition is that, not only is the full range of cases accounted for, but it achieves a greater degree of automaticity of derivation, since it makes explicit the relevance of contextual assumptions regarding speakers' capacities and interests.

In the final brief subsection, I shall look at a recent claim that relevance theory makes some wrong predictions with regard to scalar implicature.

5.3. *Less relevant but more informative?*

Matsumoto (1995) develops an account of the constraints on the first maxim of quantity (volubility, on his interpretation). He considers the possibility that a single constraint involving Sperber & Wilson's Relevance Theory might do the trick. The idea would be that a scale <S, W>, where S is the stronger item and W the weaker, licenses a scalar implicature only if the following condition is met:

(74) *Relevance condition on scalar implicatures:*
The use of W instead of S must not be attributed to S being less relevant in Sperber & Wilson's sense (i.e., carrying fewer contextual effects and/or requiring more processing effort).

(Matsumoto 1995, 53)

That is, the hearer of an utterance containing a weaker element from a scale is entitled to assume maximal informativeness (and so derive a scalar implicature), unless the use of the stronger (more informative) element from that scale would carry fewer contextual effects and/or require more processing effort. Matsumoto considers this inferior to what he calls "the conversational condition on Horn scales" (see section 2.3 above).

His allegations are directed at the 1986 version of the presumption of optimal relevance but, as far as I can tell, would apply equally to the revised version. First, it's important to recognize the oddity of what he is doing here: it is an attempt to combine a volubility principle with (part of) relevance theory, giving the latter precedence over the former. Since the communicative principle of relevance is intended to replace the whole Gricean system, including both the quantity maxims and the maxim of relation, this endeavour is bound to give rise to distortions, if not downright contradictions, which I shan't try to unravel here. Still, I think it is of some interest to consider the sort of counterexample Matsumoto is trying to construct, since, as far as I can see, it can't be done.

His key counter-example is the following, where an utterance of (75a) would most likely give rise to the scalar implicature in (75b), and this, according to Matsumoto, is contrary to the predictions of the relevance condition:

(75) a. It was a little bit more than warm yesterday, and it is just plain hot today.
b. (The speaker believes) it was not hot yesterday.

His (overly) swift explanation of the problem here is that the stronger expression "(just plain) hot" presumably requires less processing effort than the weaker expression "a little bit more than warm" since it is both briefer and more frequent. So relevance considerations predict that the implicature should not arise; since it does, relevance theory gets it wrong. Unfortunately, he has got his wires crossed here. If we grant that the stronger expression is less effort demanding than the weaker one, then it is more relevant than the weaker one (assuming no appreciable difference between them in contextual effects); so the fact that the speaker did not make the stronger statement "it was plain hot yesterday" canNOT be explained on the basis that it would have involved more processing effort and would, therefore, have been less relevant (other things being equal). In short, the relevance condition is met and so it does not block the occurrence of the scalar implicature.

What Matsumoto is looking for is a case where the stronger expression clearly requires more processing effort than the weaker one, making it possible that it is for that reason that the speaker chose W rather than S. Consider the following:

(76) a. It was warm yesterday, and it's verging on the unbearably hot today.
b. (The speaker believes) it was not verging on the unbearably hot yesterday.

Assuming an utterance of (76a) does implicate (76b), this would appear to be the sort of counterexample Matsumoto had in mind: it seems that S ("verging on the unbearably hot") would require more processing effort than W ("warm"), hence the use of W instead of S can be attributed to S requiring more processing effort, and so being, in this respect, less relevant than W; this should stop the scalar implicature going through. Well, let's consider this in realistic left-to-right processing terms. First, take it as a response to the question "Are you having hot weather?". At the end of the first clause the hearer would have derived the implicature "(The speaker believes) it wasn't hot yesterday". Would that then be altered to "It wasn't verging on the unbearably hot yesterday" as a result of processing the second clause? This seems very unlikely. Let's consider it now with no preceding utterance: by the point at which the first clause has been processed there might well be no scalar implicature at all, but, certainly, there would not be the implicature "It wasn't verging on the ... etc." which is the one at issue here, containing as it does the long effort-requiring term. It wouldn't be there precisely because of its inaccessibility. Supposing it does arise once the second clause has been processed, why does it? The answer, obviously, is that the phrase "verging on the unbearably hot" is explicitly encoded in the second clause and so made as

highly accessible as a concept can be made. In other words, if that implicature regarding yesterday's weather is derived at this point (at the end of the second clause), it is precisely because the S term is highly accessible and its 'incorporation', as it were, into an implicature takes minimal effort. So, in fact, this turns out not to be a counterexample to the relevance condition (74) either.

Finally, let's change the example so that the lengthy strong term comes before the briefer weak term in the left-to-right processing of the utterance. Again, the idea is that an utterance of (77a) would most likely implicate (77b):

(77) a. It was verging on the unbearably hot yesterday; today it's warm.
b. It's not verging on the unbearably hot today.

This would be accounted for in essentially the same way as (76), the only difference being that the clause from which the implicature is inferred (the second one) would render it immediately, as it were, because the preceding clause has made the stronger concept "verging on the unbearably hot" highly accessible. I do not see how to construct an example which has the properties Matsumoto is looking for; that is, where the stronger term takes appreciably more processing effort than the weaker term (hence is less relevant) and yet the scalar implicature (that S is not the case, as far as the speaker knows) goes through. If this is right, it looks like a nice piece of evidence in favour of Relevance Theory.

6. Conclusion

The main points I set out to make in this paper are the following:
(a) The two central informativeness principles of the neo-Griceans, specifically of Horn and Levinson, give rise to essentially the same result: a strengthening or narrowing down of the encoded meaning of the utterance.
(b) A relevance principle is needed to constrain both the Q-principle, which generates scalar implicatures, and the R/I principle, which licenses unlimited informational enrichment along uncontroversial lines.
(c) Among scalar terms, the cardinals are special; their semantics is either punctual or does not specify any one of the three interpretations, "at least n", "at most n", "exactly n", and their ultimate, pragmatically determined, interpretation contributes to the proposition expressed by the utterance (hence to its truth-conditions).
(d) The relevance-theoretic approach to utterance interpretation employs a single communicative principle, which some have seen as too reductive and unlikely to have the necessary detailed explanatory power. The capacity of the theory to account for the derivation of a wide range of implicatures shows this worry to be misplaced.
(e) The communicative principle of relevance based on the presumption of optimal relevance (1996/95) is more adequate than any of the quantity-based systems, in that its predictions are more accurate and, for those cases which they can both handle, its derivations are smoother.

NOTES

* This paper was not in fact given at the Osaka conference, since it was decided that the

meeting should kick off with a general introductory overview of the theory, which I gave. So this is a new paper and some of the references postdate the conference by several years. Many thanks to Deirdre Wilson and Vladimir Zegarac for highly relevant discussion of some of the issues here.

1. I distinguish the terms "post-Gricean" and "neo-Gricean". Post-Gricean refers to all those approaches to pragmatics that take the Gricean inferential approach to communication as their starting point and so includes relevance theory. By neo-Gricean I mean those approaches that function with some version or other of the original Gricean maxims and the Cooperative Principle; relevance theory, of course, stands outside this category.

2. Truthfulness maxims have been assumed to be indispensable in all systems except Relevance Theory. It is an empirically attested fact that what is said (the proposition expressed) is frequently not literally true or well evidenced, so the submaxims of truthfulness, which concern what is said, do not hold. The more general supermaxim ("Try to make your contribution one that is true") concerns the speaker's overall communicative contribution; its correct predictions are captured by the Communicative Principle of Relevance, given basic facts about cognitive processing. See Wilson (1995) for a useful discussion of these maxims and Sperber & Wilson (1995, 263-266) for a more general discussion of the place of truth in their overall cognitive theory.

3. The M-principle (which takes in two of Grice's manner maxims: "Avoid obscurity" and "Be brief") says that a state of affairs described in a marked or abnormal way is to be understood as having special features (that is, as not being a normal occurrence). It explicitly requires a transderivational process of reasoning, in that it involves the hearer in making a comparison between the expression the speaker used and another simpler, unmarked expression that the speaker could have used and which would have said essentially the same thing. I believe that the 'M-implicatures' that follow from the use of repetitions and longer or more obscure expressions fall out automatically from the comprehension strategy warranted by relevance theory, and that this is preferable to an account that involves hearers in quite effort-demanding processes of reasoning about what speakers could have said. This is touched on briefly in Carston (1990/95) and explored more fully in Carston (in preparation).

4. A number of authors have expressed doubt that there is any theoretically interesting distinction to be made between particularized and generalized implicatures: Hirschberg (1985/91, 42-44), Carston (1990/95, 229-231), Neale (1992, 524, footnote 18), Welker (1994, 21-23).

5. Blakemore (1987, 1988, 1989)'s treatment of this subset of "cue phrases", which is very different from the scalar implicature account referred to here, has set in motion an important strand of semantic work within relevance theory. She argues that these discourse connectives encode procedures (rather than concepts) whose function is not to enter into representations of the meaning of the utterance but to constrain the pragmatic inferential processes involved in deriving those representations. This idea is further developed in (Wilson & Sperber 1993a) and some of the papers in this volume involve applications of it (see the contributions of Itani, Tanaka and Yoshimura).

6. Welker (1994, 52-68) gives a very helpful overview and discussion of Atlas & Levinson (1981), Horn (1984, 1989) and Levinson (1987a). All of these develop the idea that there are two countervailing forces at work in the production and interpretation of utterances, the one enjoining volubility, the other taciturnity, each of which gives rise to a set of pragmatic inferences, though they differ in the exact nature of the principles/maxims they propose and the means by which it is determined which of the two is operational (that is, how the alleged clash is resolved) in a given instance.

7. The use by Atlas & Levinson (1981) and Levinson (1987a) of concepts of obviousness, non-controversiality and stereotypicality in characterising informational enrichments has been criticised by Carston (1990/95), (1994) and Welker (1994, 56). We both take the view that the notions are too vague and insufficiently context-specific, and that a notion of "accessible in a particular context" would be more appropriate. In Carston (1994) I argue that the concept of immediate accessibility in a context covers the Levinsonian examples and a range of further cases of informational enrichment which cannot be considered to be stereotypical or generally obvious across contexts.

8. I believe that the Communicative Principle of Relevance achieves exactly that (see Carston 1990/1995 and section 5 of this paper). Welker (1994)'s account also overcomes the problems of clash and overlap that arise in Horn's and Levinson's systems by employing a single overarching pragmatic principle, her "Revised Cooperative Principle" whose informal definition is: "Provide an utterance that: (a) brings the common ground closer to the conversational goals, and (b) is better than any other utterance you could have provided in terms of making the conversational goals true." Her main aim is to provide a formal generative account of conversational implicature for which she employs a plan-based approach and the formal resources of discourse representation theory.

9. Obviously, R&R are excluding from consideration here cases where "what is said" is not part of what is communicated; instances of metaphor and other tropes, as treated by Grice and most post-Griceans, are assumed to be such cases. Nothing hangs on this exclusion here.

10. While Atlas & Levinson (1981) and Levinson (1987a) treat the conditional perfection case as an I-implicature, others, especially recently, have argued for its being a case of scalar implicature (that is, Q-based), for instance, Koenig (1986), Matsumoto (1995, 44-51) and van der Auwera (1997). There are considerable differences among these accounts, including the particular scale of elements invoked in each case.

11. Example (37a) may be okay if *like* is given particular stress so that it can be interpreted metalinguistically. Then it would be understood along the lines of: "Neither of us could be said to have *liked* the movie - she hated it and I loved it".

12. This distribution of interpretations of number terms has also been pointed out by Campbell (1981) and Fretheim (1992). According to Fretheim, when the term occurs in the part of the utterance which is "thematized" (topicalized) by intonational means, its semantics is lower-bounded, with the possibility of an upper-bounding conversational implicature, and when it is not (that is, when it is "focalized" (that is, in the comment position)) its semantics is both lower and upper bounded (i.e. *exactly n*).

13. I do not mean to imply in these brief remarks about van Kuppevelt's and Scharten's work that their accounts of scalar inference are identical; the main difference is that van Kuppevelt sees these inferences as generated on the basis of a linguistic scale which is an ordered topic range, itself derived from the topic-forming question, while Scharten proposes a general mechanism of "exhaustive interpretation" of all material in comment position, followed by a process of "negating the complement". However, abstracting away from these differences, many interesting questions arise at the basic level of the Discourse Topic framework which they share; I cannot pursue these in detail here, but will simply mention three. First, the fact that the topic-comment difference does tend to give rise to this distinction in interpretation quite systematically is intriguing, but it is presented as an apparently arbitrary fact; it surely calls for a deeper explanation in which it follows from some more fundamental principles at work in interpretation. Second, although both van Kuppevelt and Scharten consider the upper-bounded interpretation to be a matter of semantics, it is a default defeasible inference (Scharten herself gives a list of what she calls 'escapes from exhaustive interpretation' pp.104-109), and so would seem inherently pragmatic on most people's understanding of the semantics/pragmatics distinction. Neither of these deeper theoretical issues is addressed by van Kuppevelt or Scharten. Finally, there are some clear counter-examples to the thesis, for instance:

Q: How many months} have 28 days?
Which months }
A1: **One - February**.
A2: **They all** do.

(comment in bold)

The response in A1 involves treating the topic-forming question as asking which/how many months have exactly 28 days, while the response in A2 takes it to have asked how many months have at least 28 days. Van Kuppevelt's account would predict the A2 response but, in fact, the great majority of responses are of the A1 type and when the A2 response is pointed out people feel that they have been tricked. The reason seems obvious; hearers are doing their best to interpret the question as relevant (compare the question "How many weeks have six days?" which is likely to provoke a "what a dumb question!" sort of response). Considerations of relevance (pointfulness) cannot be ignored and it seems very likely that the certainly strong tendency for number terms in comment position to be interpreted as *exactly n* and in topic position as *at least n* is ultimately explainable in these terms too.

14. Scharten is committed to Seuren (1985)'s Discourse Semantics framework, which is a modern day generativist version of the classic code model approach to verbal communication and utterance meaning, which has developed out of the generative semantics tradition. This approach ignores (or is unconvinced by) the Gricean inferential model of communication, the semantic underdeterminacy thesis and the evidence for the acutely context sensitive nature of utterance interpretation (see, for instance, Sperber & Wilson 1986/95, Carston 1988 and forthcoming) and much current work in psychology on human cognitive processing, in particular, our almost reflex-like capacity for attributing complex mental states to each other (see, for instance, Sperber (1994) and the papers in Whiten (1991)).

15. As well as supporting my arguments for a sense-general semantics for number terms, Atlas (1990, 1992) gives arguments for a further claim: that the semantics of natural language number terms should be distinguished from the corresponding numerals, 1, 2, 3, etc., which are the names of natural numbers. I think he is probably right about this.

16. It's notable that Grice gave only this one example of a maxim clash. It looks unlikely to me that clashes among his other maxims could arise with the possible exception of some of the manner maxims among themselves (one might choose a longer expression, e.g. *salt and pepper*, in order to avoid a more obscure one, e.g. *condiments*, or vice versa). The neo-Gricean finding of a clash between the two quantity maxims seems to be based on a misinterpretation of the first maxim of quantity as a maxim enjoining maximal informativeness (see Green (1995)).

17. Welker (1994, 77-79), in the context of a very interesting discussion of relevance theory (necessarily involving the unrevised presumption of relevance), considers the theory's predictions of the occurrence and non-occurrence of scalar implicatures. She finds that relevance theory correctly predicts the nonoccurrence of a scalar implicature in, for example, the following cases:

    (i) Context: If all the cookies are gone, then A must bake more.
    B (to A): ?? John ate some of the cookies.

    (ii) Context: If John ate some of the cookies, A will punish him.
    B (to A): John ate some of the cookies.

    In (i), B's utterance is correctly predicted as irrelevant in the given context (likely to elicit a "so what?" reaction), and in (ii), B's utterance is optimally relevant (has sufficient effects) without the scalar implicature "John didn't eat all the cookies". She is sceptical, however, about the capacity of the theory (on a literal interpretation of the 1986 formulation) to predict that a scalar implicature will occur in some instances where it does; this chimes with Sperber & Wilson (1995)'s reflections on example (70).

18. These two epistemic possibilities have long been noted by neo-Griceans and they often derive scalar implicatures in two stages, first the weaker "speaker doesn't know whether ..." and then a strengthening to "speaker knows/believes that not ...." (see, for instance, Harnish (1976, 353), Levinson (1983, 134-5), and Horn (1989, 233-344); in Hirschberg (1985/91)'s formalisation only the weaker type is derived). Interestingly, though, the parallel between the first stage here and the "South of France" case is seldom noted; of course, the stronger type of quantity implicature does not arise for this example because there is no more specific piece of information up for consideration (there is no scale available), though the example could be set up differently so that there was and so that this would feature in an implicature: "The speaker believes that Pierre does not live in Aix-en-Provence". Given the parallel, it would seem to follow that the first stage in the neo-Gricean accounts must involve a clash between their quantity (= volubility) principle and the second truthfulness maxim enjoining the making of only evidenced statements. This point is not generally recognised, with the interesting exception of Matsumoto (1995, 23-24) who derives the weaker cases in exactly the same way as the South of France example and the stronger cases as a clash between quantity-1 and the *first* maxim of truthfulness ("Do not say what you believe to be false").

19. This sort of example was also discussed by O'Hair (1969) in an insightful and prescient early attempt to find an adequate informativeness principle. After trying out various formulations and testing their predictions against a range of cases where negative implicatures concerning stronger propositions than the one expressed either clearly did or did not arise, he ended up with the following principle:

> "Unless there are outweighing good reasons to the contrary, one should not make a weaker statement rather than a stronger one if the audience is interested in the extra information that would be conveyed by the latter."

It is clear from his discussion that "outweighing good reasons to the contrary" include other interests and commitments that a speaker might have, which conflict with giving the audience the higher level of information. This principle bears significant resemblance to the relevance-theoretic principle, but it lacks the cognitive basis of the latter and so misses the crucial point that intrinsic limitations on audiences' cognitive processing resources entail that considerations of the effort involved in utterance understanding have to be accommodated by any realistic pragmatic principle.

20. In the neo-Gricean nomenclature of implicature, those adverted to here (as not occurring) are called "clausal" rather than "scalar"; the idea is that if the speaker asserts a complex expression, say "P or Q", which contains an embedded sentence or sentences which it does not entail, here both "P" and "Q", and there is an alternative expression of roughly equal brevity which contains the same embedded sentence(s) and does entail it/them, here "P and Q" or just "P" or just "Q", then the speaker implicates that she doesn't know whether the embedded sentence(s) is/are true or false. Like scalar implicatures, these clausal implicatures are taken to be generated by the first maxim of quantity or Q-principle (see Gazdar 1979, 59-61; Levinson 1983, 136-7).

REFERENCES:

Atlas, J.

1989 *Philosophy Without Ambiguity*. Oxford: Oxford University Press.

1990 "Implicature and logical form: The semantics-pragmatics interface", Lecture 3. Manuscript of lecture delivered at Second European Summer School in Language, Logic and Information. Katholieke Universiteit Leuven.

1992 "Why 'three' does not mean 3: scalar implicatures, truth- conditions, and meaning." Unpublished ms. Pomona.

Atlas, J. & Levinson, S.

1981 "It-clefts, informativeness, and logical form: radical pragmatics (revised standard version)." In P. Cole (ed.), *Radical Pragmatics*, 1-61. New York: Academic Press.

Blakemore, D.

1987 *Semantic Constraints on Relevance*. Oxford: Blackwell.

1988 "The organisation of discourse." In F. Newmeyer (ed.), *Linguistics: The Cambridge Survey* vol.IV, 229-50. Cambridge: Cambridge University Press.

1989 "Denial and contrast: A relevance-theoretic analysis of *but*." *Linguistics and Philosophy* 12, 15-37.

Breheny, R.

1997 "A unitary approach to the weak and strong interpretation of definites." *UCL Working Papers in Linguistics* 9.

Campbell, R.

1981 "Language acquisition, psychological dualism and the definition of pragmatics." In H. Parret et al. (eds.), *Possibilities and Limitations of Pragmatics*, 93-103. Amsterdam: John Benjamins.

Carston, R.

1985 "A reanalysis of some 'quantity implicatures'." Unpublished ms., University College London.

1988 "Implicature, explicature and truth-theoretic semantics." In R. Kempson (ed.), *Mental Representations: the Interface between Language and Reality*, 155-181. Cambridge: Cambridge University Press. Reprinted in S. Davis (ed.) 1991. *Pragmatics: A Reader*, 33-51. Oxford: Oxford University Press.

1990 "Quantity maxims and generalised implicature." *UCL Working Papers in Linguistics* 2, 1-31. Revised version (1995) in *Lingua* 96, 213-244.

1993 "Conjunction, explanation and relevance." *Lingua* 90, 27-48.

1994 "Conjunction and pragmatic effects." In R. Asher (ed.), *Encyclopedia of Language and Linguistics* vol. 2, 692-698. Pergamon Press and Aberdeen University Press.

1996 "Metalinguistic negation and echoic use." *Journal of Pragmatics* 25, 309-330.

(forthcoming) *Pragmatics and the Explicit/Implicit Distinction*. PhD thesis, University of London. To be published by Blackwells.

(in preparation) “Relevance, processing effort and 'manner' implicatures."

Fauconnier, G.

1975 "Pragmatic scales and logical structures." *Linguistic Inquiry* 6, 353-75.

Fretheim, T.

1992 "The effect of intonation on a type of scalar implicature." *Journal of Pragmatics* 18, 1-30.

Gazdar, G.
1979 *Pragmatics: Implicature, Presupposition, and Logical Form.* New York: Academic Press.

Green, M.
1995 "Quantity, volubility, and some varieties of discourse." *Linguistics and Philosophy* 18, 83-112.

Grice, H.P.
1961 "The causal theory of perception." *Proceedings of the Aristotelian Society*, supplementary volume 35, 121-168. Reprinted in Grice 1989, 224-247.

1967 "Logic and conversation." *William James lectures*. Printed in Grice 1989, 1-143.

1975 "Logic and conversation." In P. Cole and J. Morgan (eds.), *Syntax and Semantics 3: Speech Acts*, 41-58. New York: Academic Press. Reprinted in Grice 1989, 22-40.

1978 "Further notes on logic and conversation." In P. Cole (ed.), *Syntax and Semantics 9: Pragmatics*, 113-127. New York: Academic Press. Reprinted in Grice 1989, 41-57.

1989 *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Gundel, J., Hedberg, N. & Zacharski, R.
1990 "Givenness, implicature and the form of referring expressions in discourse." *Proceedings of the 16th Annual meeting of the Berkeley Linguistics Society, Parasession on the Legacy of Grice*, 442-453.
1993 "Cognitive status and the form of referring expressions in discourse." *Language* 69, 274-307.

Harnish, R.
1976 "Logical form and implicature." In T. Bever et al. (eds.), *An Integrated Theory of Linguistic Ability*, 313-91. New York: Harvester Press.

Hirschberg, J.
1985 *A Theory of Scalar Implicature*. PhD dissertation. University of Pennsylvania. Published 1991 in the series *Outstanding Dissertations in Linguistics*. New York: Garland.

Horn, L.
1972 *On the Semantic Properties of Logical Operators in english*. PhD dissertation, University of California, LA.

1984 "A new taxonomy for pragmatic inference: Q-based and R-based implicature." In D. Schiffrin (ed.), *Meaning, Form and Use in Context (GURT '84)*, 11-42. Washington: Georgetown University Press.

1985 "Metalinguistic negation and pragmatic ambiguity." *Language* 61, 121-74.

1989 *A Natural History of Negation*. Chicago: University of Chicago Press.

1992 "The said and the unsaid." *Ohio State University Working Papers in Linguistics (SALT II Proceedings)* 40, 163-192.

1996 Presupposition and implicature. In S. Lappin (ed.), *The Handbook of Contemporary Semantic Theory*, 299-319. Oxford: Blackwell.

Iwanska, L.

1996 "Toward a formal account of context-dependency and underspecificity of natural language." Paper given at AAAI Symposium on Computational Implicature, March 1996.

Keenan, E.

1976 "The universality of conversational postulates." *Language and Society* 5, 67-80.

Kempson, R.

1986 "Ambiguity and the semantics-pragmatics distinction." In C. Travis (ed.),*Meaning and Interpretation*, 77-103. Oxford: Blackwell.

Koenig, E.

1986 "Conditionals, concessive conditionals and concessives." In E. Traugott et al. (eds.) *On Conditionals*, 229-246. Cambridge: Cambridge University Press.

Koenig, J.

1991 "Scalar predicates and negation: punctual semantics and interval interpretations." *Chicago Linguistic Society 27, Part 2: Parasession on Negation*, 140-155.

Levinson, S.

1983 *Pragmatics*. Cambridge: Cambridge University Press.

1987a "Minimization and conversational inference." In J. Verschueren and M. Bertuccelli-Papi (eds.), *The Pragmatic Perspective*, 61-129. Amsterdam: John Benjamins.

1987b "Pragmatics and the grammar of anaphora." *Journal of Linguistics* 23, 379-434.

1988 "Generalized conversational implicature and the semantics/ pragmatics interface." Unpublished ms.

1989 "A review of Relevance." *Journal of Linguistics* 25, 455-472.

1991 "Pragmatic reduction of the Binding Conditions revisited." *Journal of Linguistics* 27, 107-161.

(forthcoming) *Presumptive Meanings: The Theory of Generalized Conversational*

Implicature. Cambridge: Cambridge University Press.

Matsumoto, Y.
1995 "The conversational condition on Horn scales." *Linguistics and Philosophy* 18, 21-60.

Neale, S.
1992 "Paul Grice and the philosophy of language." *Linguistics and Philosophy* 15, 509-559.

Oberlander, J. & Knott, A.
1996 "Issues in cue phrase implicature." Paper given at the AAAI-96 Spring Symposium on Computational Implicature, Stanford, March 1996.

O'Hair, S.
1969 "Implications and meaning." *Theoria* 35, 38-54.

Recanati, F.
1989 "The pragmatics of what is said." *Mind and Language* 4, 295-329.

Richardson, J.F. and Richardson, A.W.
1990 "On predicting pragmatic relations." *Proceedings of the 16th Annual meeting of the Berkeley Linguistics Society, Parasession on the Legacy of Grice*, 498-508.

Roberts, C.
1996 "Information structure, plans and implicature." Talk given in the AAAI-96 Spring Symposium Series: Computational Implicature: Computational Approaches to Interpreting and Generating Conversational Implicature.

Sadock, J.
1984 "Whither radical pragmatics?" In D. Schiffrin (ed.), *Meaning, Form and Use in Context: Linguistic Applications*. Georgetown University Roundtable. Washington: Georgetown University Press.

Scharten, R.
1997 *Exhaustive Interpretation: A Discourse-Semantic Account*. PhD thesis, Katholieke Universiteit Nijmegen.

Seuren, P.
1985 *Discourse Semantics.* Oxford: Blackwell.

1993 "Why does 2 mean '2'? Grist to the anti-Grice mill." In E. Hajicova (ed.), *Functional Description of Language: Proceedings of the Conference* (Prague, Nov. 24-27, 1992), 225-235. Faculty of Mathematics and Physics, Charles University, Prague.

Sperber, D.
1994 "Understanding verbal understanding." In J. Khalfa (ed.), *What is Intelligence?*,

179-198. Cambridge: Cambridge University Press.

Sperber, D. and Wilson, D.

1986 *Relevance: Communication and Cognition*. Oxford: Blackwell.

1987 "Precis of Relevance: communication and Cognition, and open peer commentary." *Behavioral and Brain Sciences* 10, 697-754.

1995 "Postface." In D. Sperber and D. Wilson, *Relevance: Communication and Cognition*, second edition. Oxford: Blackwell.

Strawson, P.

1952 *Introduction to Logical Theory*. London: Methuen.

1964 "Identifying reference and truth-value." *Theoria* 30, 96-118. Reprinted in Strawson, P. (ed.), 1971. *Logico-Linguistic Papers*, 75-95. London: Methuen & Co.

Travis, C.

1985 "On what is strictly speaking true." *Canadian Journal of Philosophy* 15: 187-229.

van der Auwera, J.

1997 "Conditional perfection." In A. Athanasiadou and R. Dirven (eds.), *On Conditionals Again*, 169-190. Amsterdam: John Benjamins.

van Kuppevelt, J.

1991 *Topic en comment. Expliciete en Impliciete Vraagstelling in Discourse*. PhD thesis, Katholieke Universiteit Nijmegen.

1995 "Discourse structure, topicality and questioning." *Journal of Linguistics* 31, 109-147.

1996a "Inferring from topics: scalar implicatures as topic-dependent inferences." *Linguistics and Philosophy* 19, 393-443.

1996b "In defense of semantics: scalar inferences as topic-dependent entailments." In B. Di Eugenio et al. (eds.), *AAAI Spring symposium on Computational Implicature: Computational Approaches to Interpreting and Generating Conversational Implicature*, Stanford University Press.

Verkuyl, H. and van der Does, J.

1995 "The semantics of plural noun phrases." In J. van der Does, J. and J. van Eijk (eds.), *Quantifiers, Logic, and Language*, 337-374.

Wainer, J. and Maida, A.

1990 "Good and bad news in formalizing generalized implicatures." *Proceedings of the 16th Annual meeting of the Berkeley Linguistics Society, Parasession on the Legacy of Grice*, 530-540.

Welker, K.

1994 *Plans in the common ground: toward a generative account of implicature*. PhD thesis, Linguistics department, The Ohio State University.

Whiten, A. (ed.)

1991 *Natural Theories of Mind*. Oxford: Blackwell.

Wilson, D.

1995 "Is there a maxim of truthfulness?" *UCL Working Papers in Linguistics* 7, 197-212.

Wilson, D. and Sperber, D.

1993a "Linguistic form and relevance." *Lingua* 90, 1-25.

1993b "Pragmatics and time." *UCL Working Papers in Linguistics* 5, 277-298. Revised version in this volume.