Scalar Implicatures and Their Interface with Grammar.

Gennaro Chierchia
3: 245-264.
1. INTRODUCTION

Scalar implicatures (SIs) are exemplified as follows:

(1a) John: What can I bring to the dinner party? Wine and dessert? 
Mary:  
  i. Bring wine or dessert.  
  ii. Don’t bring both.

(1b) John: Are you done with grading the assignments? 
Mary:  
  i. I have graded some.  
  ii. I haven’t graded all.

Plausibly, Mary in example 1a,i intends to convey to her guest, John, that he is not required or expected to bring both wine and dessert (i.e., sentence 1a,ii); in example 1b,i, she intends to convey that she graded some though not all of the assignments (i.e., sentence 1b,ii). SIs are the process whereby words like or and some, which can also have an inclusive interpretation—compatible with and and all, respectively—may come to be construed exclusively (by speakers and hearers alike). They are part of the more general phenomenon of quantity-based implicatures (QBIs). Another example of a QBI, also associated with or, is the so-called ignorance implicature:

(2a) John is at home or at the office.
(2b) (The speaker believes that) it is possible that John is at home and also possible that he is at the office.

Both SIs and ignorance implicatures have been argued to be derivable from the idea that speakers adhere to a so-called maxim of quantity, according to which they are expected to convey all the relevant information they have evidence for. Grice (1975, 1989) made this proposal while constructing a beautiful general paradigm on how core/logical interpretation comes to be pragmatically enriched. Whereas linguists generally agree that Grice's program is fundamental to understanding many forms of pragmatic enrichment, there has been considerable debate over whether SIs and other QBIs do not instead require substantially more than spelling out Grice's program, and are not in fact rooted in grammar. What is at stake is whether SIs can be viewed as stemming from general principles of rational action (the pragmatic approach) or whether they are, instead, tied to language-specific computational processes (the grammatical approach). In this article, I review the key arguments in this debate, and explore why some linguists have reached conclusions regarding SIs similar to those that Groenendijk & Stokhof (1984, pp. 368–69) reached in connection with the exhaustive interpretations of answers:

We are inclined to prefer a pragmatic strategy over the semantic one explored in this paper. Why then didn't we take this grand route over the summits of Gricean reasoning, where the air is thin, but the view so much clearer? The reason is that we don't see a pass that leads into this promised land. The informal Gricean reasoning sounds quite appealing. The problem is to make it work….
2. GRICE'S LEGACY

The reconstruction of Grice's proposal on SIs and other QBIs that one finds in semantic textbooks (e.g., Gamut 1991) can be reproduced as follows:

(3a)  
1. The speaker said ‘John or Bill will show up’  
   2. $S_s(\text{show up}(j) \lor \text{show up}(b))$  

(3b)  
1. The speaker didn’t say ‘John and Bill will show up’  
   2. $\neg S_s(\text{show up}(j) \land \text{show up}(b))$

(3c)  
1. ‘John and Bill will show up’ is relevant  
   2. $\text{show up}(j) \land \text{show up}(b) \in \text{REL}$

(3d)  
1. Say only things you believe to be true  
   2. $\forall p S_s(p) \rightarrow B_s(p)$

(3e)  
1. Say all relevant things you believe to be true  
   2. $\forall p,q \in \text{REL} \left[ p \subseteq q \land B_s(p) \rightarrow S_s(p) \right]$  
   = $\forall p,q \in \text{REL} \left[ \neg S_s(p) \rightarrow [p \not\subseteq q \lor \neg B_s(p)] \right]$

(3f)  
1. The speaker believes that John or Bill will show up  
   2. $B_s(\text{show up}(j) \lor \text{show up}(b))$

(3g)  
1. ‘John and Bill will show up’ is stronger than ‘John or Bill will show up’  
   2. $\text{show up}(j) \land \text{show up}(b) \subseteq \text{show up}(j) \lor \text{show up}(b)$

(3h)  
1. If the speaker believed that John and Bill will show up she would have said so  
   2. $\neg S_s(\text{show up}(j) \land \text{show up}(b)) \rightarrow \neg B_s(\text{show up}(j) \land \text{show up}(b))$

(3i)  
1. It is not the case that the speaker believes that John and Bill will show up  
   2. $\neg B_s(\text{show up}(j) \land \text{show up}(b))$

The derivation of the ignorance implicature is fully parallel modulo the assumption that uttering a disjunction makes the individual disjuncts relevant:
Conclusions 3i and 4i enrich the basic meaning of the uttered sentences, using maxims (like quality, quantity, and relevance) that can be regarded as spelling out a principle of least effort in communication, and pure logic. There is great beauty in this. One might wish that all forms of pragmatic enrichment could be derived in a similar way.

But, as Groenendijk & Stokhof (1984) asked, is there a pass into this promised land when it comes to SIs? Notice, in particular, that the reasoning in examples 3 and 4 afford us uniformly ignorance implicatures. The speaker is not sure that Bill will show up, nor that John will show up; and she is not sure that both will show up. This result is perfect for example 2, but it is too weak for the canonical instances of SIs such as those in example 1. We need a way of strengthening result 3i. It may seem like a small step from conclusion 3i to the exclusiveness implicature; the question is whether things are as they seem.

Note that Gricean quantity inferences can be thought of as exhaustivity clauses over the set of relevant alternatives. Such inferences are, in other words, equivalent to saying that that the only things the speaker believes among the relevant ones are those that have been uttered (the standard definition of the only operator O is given in example 5c):

\[
\begin{align*}
(4a) & = 3a \\
(4b) & = 3b \\
(4c) & = 3c \\
(4d) & = 3d \\
(4e) & = 3e \\
(4f) & = 3f \\
(4g) & = 3g \\
\end{align*}
\]
Given this equivalence, we can use $O_{\text{ALT}}(p)$ as an abbreviation for “run the Gricean reasoning on $p$, relative to ALT.”

3. PROBLEMS WITH GRICE'S LEGACY

I now turn to issues that make Grice’s program problematic for the derivation of SIs.

3.1. Opinionatedness

To obtain the desired SI from conclusion 3i, we must assume that the speaker is opinionated about the relevant alternatives, in the sense that for each alternative she either believes it to be true or believes it to be false:

$$\forall p \in \text{ALT} [B_s(p) \vee B_s(\neg p)]$$

$$B_s([\text{show up}(j) \vee \text{show up}(b)] \land \neg[\text{show up}(j) \land \text{show up}(b)]) = B(O_{\text{ALT}}(\text{show up}(j) \vee \text{show up}(b)))$$

where $\text{ALT} = \{\text{show up}(j) \vee \text{show up}(b), \text{show up}(j) \land \text{show up}(b)\}$

Through opinionatedness, we derive from example 3i that the speaker believes that either John or Bill, but not both, will show up (i.e., the desired exclusiveness implicature 6b). Opinionatedness seems like an innocent enough assumption.5

However, we do not want opinionatedness to apply to the disjuncts individually. Assuming that the speaker is opinionated in the sense of definition 6 about all the individual disjuncts would immediately lead to a contradiction. But this restriction on opinionatedness cannot be because individual disjuncts are irrelevant: If disjuncts are not relevant in uttering a disjunction, then what is? So opinionatedness cannot target all the relevant alternatives; it must target selectively the strictly scalar ones. Thus, there is something special about scalar alternatives (vis-à-vis individual disjuncts).
3.2. Maxim Suspension

A nice feature of Grice's approach is the way it accounts for the optionality of implicatures: If there are reasons to believe that a maxim (or any auxiliary principle involved in implicature computation) is suspended, then the derivation of the inference is also automatically blocked. For instance, if there are reasons to believe that the speaker has relatively little information about the relevant facts (example 7a), or if SIs are patently irrelevant to the conversational goals (example 7b), implicatures will not be computed:

(7a) I really haven’t got much of a clue as to who will show up. But John or Bill will.
(7b) I need to get into my place and forgot the keys. Is John or Bill home?

Similarly, for the maxim of quantity Grice gives the example of a treasure hunt, with the master of ceremonies uttering “The next clue is either in the garden or in the basement.” Clearly, there is every reason to think that the maxim of quantity is not being upheld in a treasure hunt context.

However, in the latter case (as with opinionatedness), things are surprisingly more complex. Grice's account leads us to expect that as soon as the maxim of quantity is suspended all quantity implicatures, including the scalar ones, should be blocked, because a premise essential to all QBIs fails to hold. Yet this is not quite what happens. Fox (2014) constructs a simple thought experiment showing that SIs (but not ignorance implicatures) can still be drawn even when the maxim of quantity cannot possibly be operative. Imagine a TV show in which money is hidden in several boxes of a set A, B, C, and so on, and the host is supposed to provide clues to participants as to where. In such a situation, the host might utter sentence 8a or 8b, even if she fully knows where the money is:

(8a) Host: There is money in box A or in box B.
(8b) Host: There is money in box A or box B or both.

If the host says sentence 8a, the participant might afterward raise objection 9a.

(9a) Participant: What you said was wrong. You said there was money in box A or in box B. In fact, there was money in both.
(9b) Participant: #What you said was wrong. You said there was money in box A or in box B or both. In fact, there was money in both.

Reaction 9a seems perfectly natural and legitimate (and contrasts with reaction 9b). Yet this is a minimal variant of Grice's own illustration of how the maxim of quantity can be called off (see Fox 2014 for a detailed discussion), and without such a maxim we should be unable to draw SIs. Thus, there is a second way in which SIs appear to be special and surprising among implicatures. SIs may persist even when the maxim of quantity is not operative.
3.3. Relevance

How do we determine what is a relevant alternative? The simplest answer to this daunting question might be that the context (the speaker's communicative intention, the information available to the illocutionary agents, etc.) guides us toward what matters. But, as illustrated above, things are not quite as free as that. Scalar alternatives (as opposed to, e.g., individual disjuncts) are selectively targeted by opinionatedness; moreover, scalar alternatives are not automatically deactivated when the maxim of quantity is called off. On the basis of considerations that overlap those discussed in Sections 3.1 and 3.2, Horn (1989) formulated the hypothesis that Gricean enrichments may have a privileged relationship to what have come to be known as Horn scales, namely paradigmatic variants of quantity expressions of the following kind:

\[(10a) \quad \text{and, or} \]
\[(10b) \quad \text{some, many, most, every} \]
\[(10c) \quad \text{can, must, etc.} \]

If we assume that the Gricean reasoning works as in reconstructions 3 and 4 plus opinionatedness, and that Horn scales provide a default setting for what alternatives are to be taken as relevant, we arrive at SIs in a way that still closely resembles Grice's original proposal. This is more or less what has come to be known as the neo-Gricean approach (e.g., Horn 2004).

Although the neo-Gricean line of research has been quite fruitful, linguists have known for a long time that SIs are not confined to paradigmatic Horn scales. Particularly important in this discussion has been the following pattern, originally put forth by Matsumoto (1995):

\[(11a) \quad \text{Yesterday it was warm.} \]
\[(11b) \quad \text{Today it is a little bit more than warm.} \]
\[(11c) \quad \text{Yesterday it wasn’t a little bit more than warm.} \]
\[(11d) \quad \text{ALT = \{warm, a little bit more than warm\}} \]

Whereas sentence 11a in isolation might not have a particularly strong SI, when followed by sentence 11b it robustly triggers SI 11c. This observation prompts three conclusions. First, SI is not limited to Horn scales. Second, the context may bring scales to salience. And third, syntactically complex items (a little bit more than warm) may qualify, through contextual salience, as scalemates to morphologically simple items (warm).

With that in mind, consider example 12:

\[(12a) \quad \text{I graded some of my assignments. Bill graded many or all of his.} \]
\[(12b) \quad \text{I graded some of my assignments. Bill graded just some of his.} \]
Example 12a is natural enough. The complex determiner many or all in the second sentence seems able to become a contextual alternative to some in the first sentence, forming the scale \{some, many or all\} as one would expect given Matsumoto's observation, and prompts a not many SI. Example 12b, by contrast, is strange. Why? On the basis of the neo-Gricean approach, it is not obvious. In fact, example 12b should have a specific implicature. The some in the first sentence of example 12b should have its basic meaning (compatible with all), and it should form a contextual scale \{some, just some\} with the complex determiner in the second sentence, which should therefore trigger a not just some implicature. Thus, example 12b should ultimately mean something like “I graded some, though not just some, of my assignments, while Bill graded just some of his.” In other words, the first sentence should be a roundabout way of saying “I graded all of my assignments.” But it isn't.

Here is another way of stating the problem. Consider or again. A sentence like p or q could have as an alternative p or q but not both, for that would be relevant, and can be explicitly made salient in the context. If we run the Gricean reasoning over p or q using as an alternative p or q but not both, p or q would implicate p and q. This never happens. It is not clear why, on the neo-Gricean approach. The reason is not that complex scalar expressions cannot count as alternatives to simple ones, for this typically happens in context. Yet no amount of contextual priming can create scales like \{some, just some\} or \{or, or but not both\}. At a descriptive level, the constraint seems to be that alternatives must have the same monotonicity. All Horn scales do; so does the scale \{warm, a little bit more than warm\}. But the scale \{some, just some\} does not have a uniform monotonicity: some is upward monotone; just some is nonmonotone.

The outcome of this discussion is as follows. Horn's idea of having formal constraints on alternatives is a step in the right direction. But (a) it conceptually changes the Gricean picture by linking it to a formal/grammatical constraint on scales, and (b) it does not explain how the context can affect the set of formally defined alternatives. What is clear is that an appeal to relevance per se does not seem to be of great help.

### 3.4. Embedding

Embedding presents a further problem for the (neo-)Gricean approach. It seems that SIs can be embedded, as demonstrated by downward-entailing (DE) contexts, like the antecedents of conditionals:

\[(13a) \quad \text{If some students in your class are having difficulties, talk to them}
\]
\[(13b) \quad \text{If you have cheese or dessert, you'll be full}
\]

When a sentence like example 13a or 13b is uttered out of the blue, some and or are not construed exclusively. For instance, sentence 13a is not typically understood as meaning “If some though not all of the students are having difficulties, talk to them.” However, examples 13a and 13b could continue as follows:

\[(14a) \quad \ldots \text{but if all of them are having difficulties, don’t talk to them; talk to me first.}
\]
\[(14b) \quad \ldots \text{but if you have cheese and dessert, you’ll get sick.}
\]
These continuations force hearers to essentially embed an exclusiveness implicature in the antecedent of a conditional. And in planning an utterance like sentence 13a with continuation 14a, the speaker relies on the rhetorical effect that continuation 14a will have.

The existence of embedded implicatures of this sort is not readily consistent with the Gricean picture. The way in which implicatures are derived in that framework is by reasoning about the utterance as a whole, the speaker's intentions, and so forth. It is supposed to leave the compositional system intact. And yet the way one would want to interpret sentences 13a and 13b in the context of continuations 14a and 14b is by treating the antecedents of sentences 13a and 13b as if they were autonomous utterances, and running the Gricean reasoning on them. The result would be as follows:

(15a) \[ \text{if } O_{\text{ALT}}(\text{some students in your class...}), \text{ talk to me, but...} \]
(15b) \[ \text{if } O_{\text{ALT}}(\text{you have cheese or dessert}), \text{ you will be full, but...} \]

[Recall that $O_{\text{ALT}}(p)$ is short for “run the Gricean reasoning on $p$”.] These observations seem to suggest that one of the main tenets of the Gricean paradigm, namely that the compositional system is “encapsulated” from pragmatic enrichments, has to be weakened.

A similar point can be made in connection with contexts that are non-DE, like the so-called Hurford disjunctions. It is well known that a disjunction is infelicitous if one of the disjuncts entails the other (example 16a). The only known exception is constituted by sentences with scalar items (example 16b, discussed in Gazdar 1979):

(16a) i. #John either ate or ate pizza.
     ii. #John is either married or married to Sue.
(16b) i. John spoke either to Sue or Bill or to both.
     ii. Either John graded some of his assignments, or he graded them all.

The deviance of sentence 16a receives a natural account on the assumption that whenever a disjunction entails another, the former becomes redundant. In contrast, the naturalness of sentence 16b can be explained in a straightforward manner by assuming that, for instance, the first pair of disjuncts in sentence 16b,i, namely (John spoke to) Sue or Bill, is construed exclusively, and therefore the first disjunction does not entail (John spoke to) Sue and Bill. The same is true for sentence 16b,ii. But if exclusive interpretations of scalar items come from implicatures, this means that implicatures must be added at an embedded level (and, in fact, at the level of subsentential units).

### 3.5. Obligatoriness

QBIs can be canceled, but the ease with which this happens varies. Contrast, for instance, examples 17a and 17b:
In example 17a, the speaker tries to cancel an SI; in example 17b, an ignorance implicature. The former seems more natural than the latter. To cancel ignorance implicatures, one has to resort to contexts of the so-called treasure hunt type.

There are many other cases where one may say that a certain piece of meaning is an implicature, but a noncancelable one. Consider, in particular, the meaning associated with plurality. The denotation of plural noun phrases must include singularities on the basis of constructions such as the following:

(18a) There are no cats on the mat.
(18b) If there are cats on the mat, shoo them away.

The presence of a single cat on the mat would suffice to make sentence 18a false; and if there is a single cat on the mat and the hearer doesn't shoo it away, she will have failed to comply with the request in sentence 18b. However, in non-DE contexts, the interpretation of plural noun phrases must involve more than one cat:

(19a) Those cats are noisy.
(19b) I just fed some cats.

One certainly cannot utter sentence 19a while pointing at one cat, and sentence 19b would be false (or infelicitous) if one fed only a single cat. It is tempting to analyze this phenomenon as a quantity implicature. The reason that sentences 19a and 19b must be understood as involving pluralities might be that they compete with their singular counterparts, which are in some sense stronger. For example, sentence 19b might be taken to assert that the speaker fed one or more cats and to implicate “but not just one.” Such an implicature would not arise in DE contexts such as sentences 18a and 18b because, as is well known, such contexts reverse entailment patterns; therefore, plural forms become stronger than their singular competitors, and no implicature comes about. This analysis is highly plausible and arguably explanatory. However, if semantic plurality in non-DE contexts is indeed an implicature, it is not cancelable:

(20) #I just fed some cats. In fact, I fed only one.

Descriptively speaking, the plurality implicature, which naturally “disappears” in DE contexts, seems to be obligatory in non-DE contexts. But the notion of obligatory implicature feels like an oxymoron from a Gricean perspective. Quantity implicatures rely on conversational maxims (and auxiliary assumptions like opinionatedness), and suspension of such maxims should always be possible.

3.6. Free Choice Phenomena

Sentences like those in example 21a tend to be interpreted as in example 21b:
Sentences 21a,i–ii have the core/compositional form in formula 22a but are interpreted as in paraphrases 21b,i–ii:

\[
(22a) \quad \Diamond [p \lor q]
\]
\[
(22b) \quad \Diamond p \land \Diamond q
\]

This so-called free choice (FC) effect is so robust that some researchers (e.g., Zimmerman 2000) have hypothesized that it is part of the compositional semantics. However, this seems implausible, because the FC effect can be canceled, and it disappears, by itself as it were, in DE contexts:

\[
(23a) \quad \text{You are allowed to have ice cream or cake. But I don’t remember which.}
\]
\[
(23b) \quad \text{You are allowed to have fruit. But you are not allowed to have ice cream or cake.}
\]

Example 23a seems to remove the FC interpretation from the first sentence, and the semantic interpretation of the second sentence in example 23b is of the form \[\neg \Diamond [p \lor q]\] and not of the form \[\neg [\Diamond p \land \Diamond q]\] (the latter would yield the much weaker meaning “You are not allowed to have ice cream or you are not allowed to have cake”). The FC effect extends to existentials under modals in general. Sentence 24a illustrates the effect with a modal of necessity, and sentence 24b with an indefinite:

\[
(24a) \quad \text{i. You must read this book or that book.}
\]
\[
\text{ii. FC implicature: You are allowed to read this book, and you are allowed to read that book.}
\]
\[
(24b) \quad \text{i. You may borrow a book from my collection.}
\]
\[
\text{ii. FC implicature: Every book in my collection is such that you are allowed to borrow it.}
\]

Thus, the FC interpretation of modals over disjuncts/existentials has the signature properties of QBIs: It automatically vanishes in DE contexts, and it is cancelable in upward-entailing (UE) ones. Moreover, it very much “feels like” a quantity implicature: If the speaker knew that John might be home but not at the office, she would surely have said so. Similarly, if I know that you are only allowed to have ice cream, uttering sentence 21a would be dramatically misleading. But how does one derive the FC effect from the Gricean reasoning laid out in reconstructions 3 and 4, even incremented with opinionatedness? How should the maxims be enriched or modified? What further principles should we call upon?
3.7. First En Route Summary

The Gricean approach delivers ignorance implicatures for free, so it seemed like a small step from there to SIs: appealing to opinionatedness. But closer scrutiny reveals a host of issues. First, opinionatedness must selectively target scalar alternatives over other alternatives that are clearly relevant. Second, there are cases in which the maxim of quantity is clearly suspended and yet we seem able to draw SIs. Third, the choice of alternatives seems to be context dependent but at the same time strongly constrained by grammar (perhaps, by a monotonicity constraint). Fourth, QBIs can be freely embedded. Fifth, some QBIs appear to be noncancelable. And sixth, there are robust and widespread QBIs, like the FC one, that have a Gricean “feel” but seem to fall beyond the reach of the Gricean approach.

So we face a dilemma. We can modify and enrich the original Gricean take, so as to try to derive all of these phenomena and observations. The question is what features of the original Gricean program the outcome of this modification will retain, and whether we end up sacrificing a clean pragmatics to the goal of keeping the semantics simple. Part of the appeal of Gricean pragmatics was its promise to deliver enrichments of meaning in a way that does not affect the computational/compositional semantic system of grammar and relies on general principles of rational action. This promise could still be kept when it comes to implicatures of a certain kind, for instance, “Who stole the steak? The dog looks happy.” But when it comes to quantity implicatures, the evidence reviewed in the previous section raises doubts, prima facie serious ones, on the “null hypothesis status” of any straightforward modification of the Gricean approach laid out in reconstructions 3 and 4. Again, as stated by Groenendijk & Stokhof (1984), “[t]he informal Gricean reasoning sounds quite appealing. The problem is to make it work.”

4. INVOLVING THE COMPOSITIONAL SYSTEM

The phenomenon that Groenendijk & Stokhof (1984) were concerned with was the analysis of (short) answers to questions. Questions are typically requests for complete information, and (short) answers must be construed as exhaustive relative to the answer space:11

(25a) Who of John, Bill, Mary and Sue will you invite?
(25b) i. John and Sue. I won’t invite Mary and Bill.
      ii. John or Sue. I won’t invite Mary and Bill.11

Notice that this happens even when the answer is disjunctive (and therefore partial). Answer 25b,ii clearly indicates (in the absence of explicit clues to the contrary) exclusion of the individuals who are not mentioned. It is interesting that such an answer does not necessarily require that the disjunction be exclusive, for the following continuation is possible:

(26a) Who of John, Bill, Mary and Sue will you invite?
(26b) John or Sue; possibly both. I won’t invite Mary and Bill.
Answer 26b clearly requires exclusiveness with respect to the individuals not mentioned, but not exclusiveness of the disjunction as such. This is further evidence of the highly selective character of QBIs.

On the basis of such phenomena, Groenendijk & Stokhof (1984) argue for the presence of an only-like exhaustivity operator in the semantics of answers. What naturally comes to mind, in view of the considerations in Section 3, is that their proposal might well indicate a strategy for dealing with QBIs in general. This strategy constitutes a significant shift in perspective, according to which QBIs would stem not from principles of efficient communication but rather from an exhaustivity operator present in the compositional system. The compositional system becomes somewhat more complicated, but the pragmatics may remain simple (and Gricean).

This line of inquiry can be developed in several ways, but all of the possible implementations share a number of features. First and foremost, there will be exhaustification in the semantics: Hereafter, O_{ALT}(p) no longer stands for “run the Gricean reasoning on p relative to ALT,” but instead for either a rule built into the semantic composition of specific constructions (as Groenendijk & Stokhof 1984 and Chierchia 2004 do) or an operator freely inserted at logical form (as Chierchia et al. 2011 propose). Second, the alternatives of scalar expressions are grammatically determined. For example, they can be associated with lexical items and can grow recursively, “stored” in separate semantic dimensions as on the semantics for focus (Rooth 1992). Each expression α, in other words, has an ordinary value \|α\| and a set of alternative values \|α\|_{ALT}. Alternatives provide the restriction for the operator O, just as focal alternatives do with the overt counterpart of O, namely only. In particular, scalar terms have Horn scales as their default setting, which can be modified by the context, subject to a monotonicity constraint. They also associate with domain alternatives, such as single disjuncts in the case of disjunction or subquantificational domains in the case of quantified determiner phrases like some students.

Third, optionality must be accommodated. A simple way to do so is to assume that insertion of the exhaustification operator O is free. Another way (which can coexist with the optionality of the operator) involves “pruning” (i.e., excluding) grammatically determined alternatives on the basis of contextual considerations.

The analysis sketched above provides us with a baseline approach with the following characteristics. Sentence 27a, for example, has (at least) two parses or logical forms:

(27a)  John will hire Mary or Sue.
(27b)  [John will hire Mary \lor John will hire Sue]  \text{No SI}
(27c)  O_{ALT}(J will hire M \lor J will hire S) (John will hire Mary or Sue) =  \text{SI}
        J will hire M \lor J will hire S \land \neg(J will hire M \land J will hire S)

In concrete cases, the choice between the two parses will be resolved in the same way that any other kind of ambiguity (lexical or structural) is resolved.

Exhaustification of a disjunction relative to individual disjuncts leads to contradiction:
There are a number of possible reactions to this fact. One is to say that in such a case individual disjuncts are simply ignored or pruned. This would leave ignorance implicatures to the Gricean component (which, as we saw, can derive them at no cost). Another possibility is that when a contradiction arises, one can take the speaker's beliefs into consideration and arrive at interpretations of the following form:

\[
O_{\{J\}} \text{ will hire } M, J \text{ will hire } S \} (\text{John will hire Mary or Sue}) = \\
[J \text{ will hire } M \lor J \text{ will hire } S] \land \neg J \text{ will hire } M \land \neg J \text{ will hire } S = \bot
\]

This is the ignorance implicature. The SI and the ignorance implicature can, of course, coexist:

\[
O_{\{B, J\}} \text{ will hire } M, J \text{ will hire } S \} B_{\{\} \text{ will hire } M \lor J \text{ will hire } S \} (\text{John will hire Mary or Sue}) = \\
[B_{\{\} \text{ will hire } M \lor J \text{ will hire } S} \land \neg B_{\{\} \text{ will hire } M \land \neg J \text{ will hire } S}) \land \neg B_{\{\} \text{ will hire } M \land \neg J \text{ will hire } S}]
\]

We now have an overall idea of how Groenendijk & Stokhof's (1984) proposal can be extended to QBIs in general. Before turning to the consequences of this stance with respect to the issues discussed in Section 3, I note that the complication of the compositional analysis considered here is not far fetched in view of how languages work. Linguists know that languages have phonologically unrealized elements like null pronouns, causativization operators, reflexivization ones, and so on. A case very similar to the one considered here is that of covert distributivity (discussed in Roberts 1987, among many others). A sentence like Sue and Mary ate a pizza has a distributive reading according to which Sue and Mary ate a pizza each. It seems hard to arrive at such a reading without assuming that Sue and Mary ate a pizza has as part of its logical form/semantic interpretation a null counterpart of the distributive operator each. Just as there is a null counterpart of each, there may well be a null counterpart of only.

In the remainder of this section, I go over the issues raised in Section 3 in connection with the Gricean take on quantity implicatures and see what happens to them under the new proposal.

### 4.1. Opinionatedeness

There is no need or place for opinionatedness in the present approach; rather, one freely exhaustifies alternative bearers. By a “confidence display,” a speaker can convey that an exhaustified parse is intended.

### 4.2. Maxim Suspension

Quantity implicatures are no longer driven by maxims. The treasure hunt–like context typically suspends the maxim of quantity and affects the speaker's and hearer's beliefs accordingly. But it is no longer directly expected that the suspension of the maxim of quantity simultaneously
blocks all quantity implicatures, particularly not SIs. Exhaustified parses relative to scalar alternatives are expected to be available (and possibly appropriate) in such contexts.

4.3. Relevance

The role of relevance is trivialized in favor of grammatically determined alternative sets (much as in the semantics of focus). The context can prune grammatically determined alternatives when particular alternatives lead to contradiction or when they do not affect the communicative goals. The context can also increment alternative sets by picking up complex scalar expressions from the surrounding discourse. There is, moreover, a very natural economy constraint that enables one to derive the observed monotonicity constraint on alternative sets.\textsuperscript{15} One must assume that contextual modification of alternative sets can never be such that $O_{\text{ALT}}(p) = q$, where $q$ is a member of the set of grammatically determined alternatives to $p$. For example:\textsuperscript{16}

\[
O_{\text{[John will hire Mary or Sue]}} (\text{John will hire Mary or Sue})^{16} = \left[ (\text{I will hire Mary} \vee \text{I will hire Sue}) \land \neg \text{I will hire Sue} \right] = \text{John will hire Mary.}
\]

In example 31, we exhaustify \textit{John will hire Mary or Sue}, pruning one of the disjuncts, namely \textit{John will hire Mary}. As a result, \textit{John will hire Mary or Sue} would be interpreted as \textit{John will hire Mary}, a patently crazy result. The economy condition on exhaustification (Exhaustification Economy) rules this interpretation out (because disjuncts are among the formal alternatives to disjunctions). Furthermore, it also rules out the cases discussed in Section 3.3, specifically the following:

\[
(32a) \quad ^*\text{I graded some of my assignments. John graded just some of his.}
(32b) \quad O_{\text{[some, just some]}} (\text{I graded some of my assignments}) = \text{I graded all of my assignments.}
\]

Exhaustifying the first sentence in 32\textsuperscript{a} relative to the contextually salient scalar alternatives contained in the second sentence leads the first sentence to mean \textit{I graded all of my assignments}. This interpretation is ruled out by Exhaustification Economy, as \textit{all} is a formal grammatical alternative to \textit{some}.

4.4. Embedding

Embedded SIs are clearly expected to exist on the present approach. One can freely embed null causativization operators, null \textit{each}, and so on, and there is no reason to expect null \textit{only} to have a different destiny. Embedding $O$ sometimes leads to strengthening and sometimes to weakening:
Exhaustifying the consequent of a conditional leads to strengthening (example 33c); exhaustifying the antecedent does not (example 33d). It is natural to assume that, if one is going to use O, one will tend to do so in a way that leads to addition of information (i.e., strengthening) rather than in a way that leads to loss of information (i.e., weakening):

\[
(34) \quad \text{Parsing condition on O: If using O, do so in a way that does not lead to weakening (unless weakening is necessary to avoid a contradiction).}
\]

Parsing condition 34 explains the pattern observed in Section 3.4 and repeated here:

(35a) If some students in your class are having difficulties, talk to them
(35b) … but if all of them do, do not talk to them; talk to me first.

Adding O to the antecedent of sentence 35a leads to weakening. Thus, speakers tend not to do it (and the sentence in isolation is therefore interpreted as If some or all of the students in your class…). However, when continuation 35b comes in, the “unmarked” construal of sentence 35a becomes too strong—that is, contradictory. For this reason, one inserts O in the antecedent in order to arrive at a coherent reading of the whole.

In conclusion, O can be freely embedded, modulo parsing condition 34, which accounts in a very simple way for the preferred distribution of SIs in embedded contexts.

### 4.5. Obligatoriness

The concept of obligatory implicature makes little or no sense from a Gricean perspective, but all kinds of grammatical processes can be optional or obligatory. Some verbs can have direct objects (e.g., eat); some verbs must (e.g., devour). In most constructions, adverbs are optional (John made it versus John made it easily), but in middle constructions they are obligatory (This book reads easily versus *This book reads). A theory in which QBIs are derived through a grammatical operator leads one to expect that sometimes (e.g., with certain types of alternatives) such an operator might be obligatory.

### 4.6. Free Choice

In this subsection, I attempt to explain how the FC effect might be derived as a quantity implicature in the present system, following the spirit (and almost, but not quite, the letter) of the proposal developed by Fox (2007). Take phrase 36b as an answer to request 36a:
Addressing request 36a with answer 36b conveys that nobody but John can be the committee chair. This is to be expected, as short answers are tendentially construed exhaustively. In other words, the interpretation of answer 36b would be as follows:

(36a) I want to know who of John, Mary and Sue can chair this committee.
(36b) John.
(36c) John or Sue.

This analysis follows from considering the relevant alternatives as potential answers to a question, and therefore imagining them as being, in turn, construed exhaustively. Choosing answer 36b over answer 36c (and its variants) signals that answer 36b (and its variants) is false, giving rise to the FC effect. So, in the general case, the FC reading can be obtained from individual alternatives of the form only John, as follows:

(37) John = John can be committee chair, Mary and Sue cannot = O_{\phi_j, \phi_m, \phi_s} \phi_j

It is natural to imagine that uttering disjunctive answer 36c does two things: (a) It excludes the individuals not mentioned, namely Mary, and (b) it excludes the possibility that phrase 36b is the right answer. In other words:

(38) John or Sue = John or Sue can be chair, Mary cannot, and it is also not true that only John can, nor that only Sue can = John and Sue both can, but Mary cannot.

Why does the FC effect arise only in the presence of a modal? What would happen if one tried to derive it in simple, unmodalized cases such as example 40, by choosing a suitable set of alternatives?

(40a) Mary or Sue chaired the committee.
(40b) O_{\{only M chaired, only Sue chaired\}} (M \lor S) = M \land S = Mary and Sue chaired the committee.

The consequence of exhaustifying a nonmodalized disjunction relative to only Mary chaired the committee and only Sue chaired the committee would derive the implicature that they both did. The reason that this does not and cannot happen is that it runs afoul of Exhaustification Economy: It yields an and reading, and and is surely among the grammatically determined alternatives for or.12
In sum, one can manufacture the FC reading in a simple way from individual (domain) alternatives by construing them exhaustively (i.e., as if they were addressing a question). Coupled with independently motivated principles (like Exhaustification Economy), the approach outlined in this subsection predicts the distribution of FC readings down to a remarkable level of precision (beyond what can be discussed in this review). This is a striking result. I know of no more insightful or less stipulative way of deriving FC readings.

4.7. Second En Route Summary

What were problems for the Gricean perspective seem to largely disappear or be resolved once we assume that QBIs are obtained through a covert counterpart of only and a grammatically determined set of alternatives. Generalizing what Groenendijk & Stokhof (1984) did for one particular phenomenon to all quantity enrichments of core/logical meaning seems to be an explanatory move. The problem is not whether the Gricean view can be modified so as to get similar effects, for it surely can, but rather whether this can be done with fewer stipulations and with results substantially different in empirical coverage and/or explanatory value than the result of modifying the compositional system as discussed here.

5. NEW HORIZONS

In this section, I describe new research directions opened up by the treatment of QBIs discussed in Section 4. Consider sentences 41a and 41b:

(41a)  

<table>
<thead>
<tr>
<th>Context: You have brought no books along.</th>
</tr>
</thead>
<tbody>
<tr>
<td>i. You may borrow some books from that bookshelf.</td>
</tr>
<tr>
<td>ii. You may borrow any book from that bookshelf.</td>
</tr>
</tbody>
</table>

(41b)  

| i. Borrow some books from that bookshelf! |
| ii. Borrow any books from that bookshelf! |

Modals of possibility and imperatives favor FC readings. In such contexts, “plain vanilla” indefinites, like plural some, and items that “specialize” in FC readings, like any, come to have virtually indistinguishable readings. In cases like sentence 41a.i, the hearer is allowed to borrow some books, and all the books in the speaker's library are allowable choices. The only noticeable difference between sentences 41a.i and 41a.ii seems to be that the FC effect is cancelable with some, but of course not with any:

(42a)  

You may borrow some books from that bookshelf. But only those three.

(42b)  

You may borrow any books from that bookshelf. #But only those three.

There is a strikingly simple analysis of this phenomenon. Any and some are identical: They are existential quantifiers and activate the very same set of alternatives. The FC effect is derived in exactly the same manner in sentences 41a and 41b: through a process of exhaustification fully parallel to that described in Section 4.6. The only difference between some and any is that the
alternatives associated with *some* can be ignored (i.e., contextually pruned), whereas those associated with *any* cannot. This can be taken as a lexical condition/stipulation on the meaning of *any*: “My alternatives are unprunable.” As a consequence of this lexical condition, the implicatures associated with *any* become obligatory, unavoidably limiting its distribution. In addition to its inherent simplicity, this line of analysis has the advantage of explaining why so-called FC items (FCIs) derive so often, historically, from plain vanilla indefinites (e.g., *any* has the same source as German *einige*, a plain indefinite).

Here is a second important set of facts. As discussed above, exhaustification of disjunctions relative to individual disjuncts results in contradiction, as in example 29, repeated here:

\[
(43) \quad O_{[\neg J \text{ will hire } M, \neg J \text{ will hire } S]} \text{ (John won’t hire Mary or Sue)} = \\
\neg [\text{J will hire } M \text{ or J will hire } S] = \bot
\]

Now imagine that there is a disjunctive/existential item that insists on having alternatives like those in example 43: individual alternatives. Such an item would always be contradictory in “positive” contexts, but it would be acceptable under negation:

\[
(44) \quad O_{[-J \text{ will hire } M, -J \text{ will hire } S]} \text{ (John will hire Mary or Sue)} = \\
\neg [\text{J will hire } M \text{ or J will hire } S]
\]

In DE contexts, the negative prejacent to O entails all of the alternatives (which will be also negative), and exhaustification has no effect. What remains is a disjunctive/existential item that is “happy” only in DE contexts. Does this sound familiar?

Analyses of negative polarity items (NPIs) with this exact “logic” have been explored extensively in the literature. What the current take on quantity implicatures brings to light is the properties that all of these analyses share, and what the choice points are. FCIs and NPIs are disjunctive/existential terms with wholly canonical alternatives (just like *or* and *some*). Moreover, FCIs differ minimally from NPIs in that an NPI insists on plain subdomain/individual alternatives and an FCI allows for exhaustified ones. And both FCIs and NPIs differ from plain indefinites in that the alternatives of the former cannot be ignored or pruned, which determines the different readings and distribution of FCIs and NPIs versus plain indefinites.\(^\text{18}\)

One final point: In discussing why the FC effect in English (and many other languages) requires a modal, I appealed to Exhaustification Economy. Here is the relevant example again:

\[
(45a) \quad \text{Mary or Sue chaired the committee.} \\
(45b) \quad O_{[\text{only Mary chaired, only Sue chaired}]} (M \lor S) = M \land S = \\
\text{Mary and Sue chaired the committee.}
\]

If we exhaustify a disjunctive/existential item with respect to alternatives such as those in interpretation 45b, we obtain a conjunctive/universal interpretation. In English, this interpretation is blocked by the presence of a grammatically determined alternative with that very meaning,
namely *and*. But imagine a language that lacks *and*, namely an item specifically designated for conjunction. In such a language, the blocking effect of *and* would be absent, but *or* would be free to have interpretation $45b$, and would thus oscillate between a disjunctive and a conjunctive interpretation, depending on the context, the quantity of information available to the speaker, and so on. At the same time, in negative (i.e., DE) environments, such an item would revert to a narrow-scope disjunctive interpretation.

This might be exactly what happens in languages like Walpiri ([Bowler 2014](#)) and Cheyenne ([Murray 2013](#)), which have been described as lacking a word for *and*. Also, American Sign Language (ASL) does not distinguish disjunction from conjunction and might be analyzed along similar lines ([Davidson 2013](#)). I strongly suspect that modal systems that have been reported to have a single modal auxiliary that oscillates between a ‘must’ and a ‘can’ reading, like Gitksan ([Matthewson 2013](#)), are amenable to a similar analysis. Furthermore, children allow systematically conjunctive interpretations of disjunction (e.g., [Singh et al. 2016](#)), perhaps because they are still unable to paradigmatically connect disjunction to conjunction. Thus, conjunction yields no blocking effect in child grammar, allowing a conjunctive reading of disjunction (much as it does in Walpiri or ASL). This hypothesis would also make sense of the late development of SIs in child language.

All of these new horizons are being carefully probed and tested. Whether they will ultimately pan out has yet to be determined. It is exciting, however, to see how many recalcitrant empirical facts might be brought under the same analytical lenses and perhaps receive novel, insightful explanations.

In conclusion, SIs seemed to be easy to understand by reasoning in terms of how compositional semantics becomes part of an efficient way of communicating by rational agents: the Gricean view. However, a closer look at QBI s over the past few years has exposed a number of properties that do not fit smoothly within that view. At the same time, a different take on quantity implicatures that analyzes them in terms of a covert counterpart of *only* seems to make good sense of the puzzling properties of this form of semantic enrichment. From the “alternatives and exhaustification” perspective, the universe of QBI s starts to look “normal” again, and new research horizons emerge that are capable of unifying a remarkable range of phenomena, including polarity sensitivity, FC effects, conjunctive interpretations of disjunctions in languages that lack overt conjunction, child language, and more. In other respects, the Gricean paradigm remains in place in its simplest, original form.

**disclosure statement**

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.
acknowledgments

The ideas discussed in this review are the results of endless discussions with many colleagues, especially Danny Fox and Benjamin Spector. I am also grateful to an anonymous reviewer and to Roberta Pires de Oliveira for their careful reading of a previous version.

literature cited

  - Crossref
  - Google Scholar
  - Article Location
  - More AR articles citing this reference

  - Google Scholar
  - Article Location

  - Crossref
  - Web of Science ®
  - Google Scholar
  - Article Location

  - Crossref
  - Web of Science ®
  - Google Scholar
  - Article Location

  - Google Scholar
Article Locations:

- Article Location
- Article Location
- Article Location
- Article Location
- Article Location

More AR articles citing this reference

  - Crossref
  - Google Scholar

Article Locations:

- Article Location
- Article Location
- Article Location
- Article Location
- Article Location
- Article Location

More AR articles citing this reference

  - Google Scholar

Article Locations:

- Article Location
- Article Location
- Article Location
- Article Location
- Article Location
- Article Location

- **Crnic L. 2013.** Focus particles and embedded exhaustification. *J. Semant.* 30:533–58
  - Crossref
  - Web of Science ®
  - Google Scholar

Article Locations:

- Article Location
- Article Location
• Davidson K. 2013. ‘And’ and ‘or’: general use coordination in ASL. *Semant. Pragmat.* 6:1–44
  - [Crossref](Crossref)
  - [Google Scholar](Google Scholar)

  Article Location

  - [Crossref](Crossref)
  - [Google Scholar](Google Scholar)

  Article Locations:
  - Article Location
  - Article Location
  - Article Location

  More AR articles citing this reference

  - [Crossref](Crossref)
  - [Google Scholar](Google Scholar)

  Article Locations:
  - Article Location
  - Article Location

  - [Crossref](Crossref)
  - [Web of Science ®](Web of Science ®)
  - [Google Scholar](Google Scholar)

  Article Locations:
  - Article Location
  - Article Location
  - Article Location
  - Article Location

  - [Crossref](Crossref)
  - [Google Scholar](Google Scholar)
  - [Crossref](https://doi.org/10.1007/s11054-011-9030-8)
  - [Web of Science ®](https://www.webofknowledge.com)
  - [Google Scholar](https://scholar.google.com)

  - [Google Scholar](https://scholar.google.com)

• **Gazdar G. 1979.** *Pragmatics: Implicature, Presupposition, and Logical Form.* New York: Academic
  - [Google Scholar](https://scholar.google.com)

• **Geurts B. 2011.** *Quantity Implicatures.* Cambridge, UK: Cambridge Univ. Press
  - [Google Scholar](https://scholar.google.com)

  - [Google Scholar](https://scholar.google.com)

  - [Google Scholar](https://scholar.google.com)
  - [Google Scholar](#)

More AR articles citing this reference

  - [Google Scholar](#)

More AR articles citing this reference

  - [Google Scholar](#)

More AR articles citing this reference

  - [Google Scholar](#)
  o Crossref
  o Web of Science ®
  o Google Scholar

**Article Locations:**

  o Article Location
  o Article Location

  o Google Scholar

**Article Location**

**More AR articles citing this reference**

  o Google Scholar

**Article Location**

  o Crossref
  o Google Scholar

**Article Locations:**

  o Article Location
  o Article Location

  o Crossref
  o Google Scholar

**Article Locations:**

  o Article Location
  o Article Location

**More AR articles citing this reference**
• **Magri G. 2011.** Another argument for embedded scalar implicatures based on oddness in downward entailing environments. *Semant. Pragmat.* 4:1–51
  - [Crossref](https://doi.org/10.1016/j.sempra.2011.01.001)
  - [Google Scholar](https://scholar.google.com/scholar?q=Magri+2011&btnG=&hl=en&as_sdt=0%2C5)

Article Locations:
  - [Article Location](https://doi.org/10.1016/j.sempra.2011.01.001)
  - [Article Location](https://scholar.google.com/scholar?q=Magri+2011&btnG=&hl=en&as_sdt=0%2C5)

• **Matsumoto Y. 1995.** The conversational condition on Horn scales. *Linguist. Philos.* 18:21–60
  - [Crossref](https://doi.org/10.1016/S0167-2898(95)01382-8)
  - [Web of Science ®](https://www.webofknowledge.com/)
  - [Google Scholar](https://scholar.google.com/scholar?q=Matsumoto+1995&btnG=&hl=en&as_sdt=0%2C5)

Article Location

  - [Google Scholar](https://scholar.google.com/scholar?q=Murray+SE+2013&btnG=&hl=en&as_sdt=0%2C5)

Article Location

• **Matthewson L. 2013.** Gitskan modals. *Int. J. Am. Linguist.* 79:349–94
  - [Crossref](https://doi.org/10.1353/ial.2013.0020)
  - [Web of Science ®](https://www.webofknowledge.com/)
  - [Google Scholar](https://scholar.google.com/scholar?q=Matthewson+L.+2013&btnG=&hl=en&as_sdt=0%2C5)

Article Location

More AR articles citing this reference

  - [Google Scholar](https://scholar.google.com/scholar?q=Roberts+C.+1987&btnG=&hl=en&as_sdt=0%2C5)

Article Location

  - [Crossref](https://doi.org/10.1080/01441150500478383)
  - [Web of Science ®](https://www.webofknowledge.com/)
  - [Google Scholar](https://scholar.google.com/scholar?q=van+Rooij+R,+Schulz+K.+2006&btnG=&hl=en&as_sdt=0%2C5)

Article Location
More AR articles citing this reference

  - Crossref
  - Google Scholar

  Article Location

More AR articles citing this reference

  - Google Scholar

  Article Locations:
  - Article Location
  - Article Location

  - Crossref
  - Web of Science®
  - Google Scholar

  Article Locations:
  - Article Location
  - Article Location

More AR articles citing this reference

  - Crossref
  - Google Scholar

  Article Location

  - Google Scholar

  Article Location
  - [Crossref](#)
  - [Google Scholar](#)

  [Article Location](#)

  [More AR articles citing this reference](#)

  - [Crossref](#)
  - [Google Scholar](#)

  [Article Location](#)