The Grammatical View of Scalar Implicatures and the Relationship between Semantics and Pragmatics

Gennaro Chierchia, Danny Fox, and Benjamin Spector

1. Introduction

Recently there has been a lively revival of interest in implicatures, particularly scalar implicatures. Building on the resulting literature, our main goal in the present paper is to establish an empirical generalization, namely that SIs can occur systematically and freely in arbitrarily embedded positions. We are not so much concerned with the question of whether drawing implicatures is a costly option (in terms of semantic processing, or of some other markedness measure). Nor are we specifically concerned with how implicatures come about (even though, to get going, we will have to make some specific assumptions on this matter). The focus of our discussion is testing the claim of the pervasive embeddability of SIs in just about any context, a claim that remains so far controversial. While our main goal is the establishment of an empirical generalization, if we succeed, a predominant view on the division of labor between semantics and pragmatics will have to be revised. A secondary goal of this paper is to hint at evidence that a revision is needed on independent grounds. But let us first present, in a rather impressionistic way, the reasons why a revision would be required if our main generalization on embedded SIs turns out to be correct.

In the tradition stemming from Grice (1989), implicatures are considered a wholly pragmatic phenomenon and SIs are often used as paramount examples. Within such a tradition, semantics is taken to deal with the compositional construction of sentence meaning (a term which we are using for now in a loose, non technical way), while pragmatics deals with how sentence meaning is actually put to use (i.e. enriched and possibly modified through reasoning about speakers' intentions, contextually relevant information, etc.). Simply put, on this view pragmatics takes place at the level of complete utterances and pragmatic enrichments are a root phenomenon (something that happens globally to sentences) rather than a compositional one. But if SIs can be systematically generated in embedded contexts, something in this view has got to go. Minimally, one is forced to conclude that SIs are computed compositionally on a par with other aspects of sentence meaning. But more radical task reallocations are also conceivable. While we may not be able to reach firm conclusions on this score, we think it is important to arrive at a consensus on what are the factual generalizations at stake, how they can be established, and what range of consequences they may have.

Let us rephrase our point more precisely. The semantics/pragmatics divide can usefully be lined up with compositional vs. post compositional interpretive processes. In the compositional part, basic meanings are assigned to lexical entries, which are then composed bottom up using a restricted range of semantic operations on the basis of how lexical entries are put together into phrases. For example, within current generative approaches, the basic device for syntactic composition is *merge* (a recursive binary operation that integrates constituents into larger units) and its semantic counterpart is function application ('*apply*'). Operations like *merge* and *apply* are key components of the computational system of Universal Grammar (UG); they operate in an automatic fashion, blind to external considerations, e.g., speaker intensions and relevant contextual knowledge. Sentence meaning is, thus, constructed through *apply*. But what is sentence meaning? Such a notion is often identified with truth conditions. While semantics, as we understand it, falls within this tradition, we would like to keep our options open on the exact nature of sentence meaning. For the notions of sentence content that have emerged from much recent work are way more elaborate than plain truth conditions. For example, sentence meaning has been argued to involve the computation of alternative meanings and hence to be a multidimensional phenomenon (cf. the semantics of questions, focus, etc.); or sometimes sentence meaning has been assimilated to context change potentials (cf. dynamic approaches to presuppositions and anaphora). We remain neutral here on these various options, and we do so by simply taking sentence meaning as equivalent to the output of the compositional process of interpretation as determined by UG, whatever that turns out to be.

In understanding the compositional/postcompositional divide, one further preliminary caveat must be underscored. Sentence meaning is blind to context, but not independent of it. Virutally every word or phrase in Natural Language is dependent on the context in some way or other. In particular, the meaning of sentences will contain variables and indexicals whose actual denotation will require access to factual information accessible through the context. To illustrate, consider the following standard analysis of *only* and focus association, along the lines of Rooth (1985, 1992) and Krifka (1993) (an example that will turn out to be very useful for our approach to SIs). According to Rooth a sentence like (1a) is analyzed as shown in (1b-d):

(1) a. Joe only upset [$_F$Paul and Sue]
(where [$_F$] indicates the constituent bearing focal stress)
b. only [Joe upset [$_F$Paul and Sue]]
c. $||$only [Joe upset [$_F$Paul and Sue]]$||_c = O_{ALT(D)}($ upset(john, paul + sue) =
upset(john, paul + sue) $\wedge \forall p \in$ ALT(C)[(upset(john, paul + sue)$\underline{\not\subset}$ p) $\rightarrow \neg$ p]
where p+s is the plural individual comprising Paul and Sue and '$\underline{\subset}$', '$\underline{\not\subset}$' stand for entail/does not entail respectively.
d. ALT(D) = { john upset u: u$\in$D} =
{ joe upset lee, joe upset sue, joe upset kim,…}

Something like (1a) has the Logical Form in (1b), where *only* is construed as a sentential operator, and is interpreted as in (c). Such interpretation, informally stated, says that Joe saw Paul and Sue and that every member of the contextually restricted set of alternatives ALT not entailed by the assertion must be false. Thus, in particular, *Joe saw Paul* is entailed by the assertion, and hence has to be true, but *Joe saw Kim* is not, and hence must be false. The set ALT is specified as in (1d). Such a set is generated by UG driven principles through a separate recursive computation (and this is part of what makes sentence meaning multidimensional). In (1c), there is one variable whose value has to be picked up by pragmatic means: D, the quantificational domain. The determination of D's value is a pragmatic, 'postcompositional' process.[1]

So, pragmatics, as understood here, is the process whereby speakers converge on reasonable candidates as to what the quantificational domain may be; it is also the process whereby a sentence like (1a) may wind up conveying that the meeting was a success (because, say, Joe managed to keep the number of upset people to a minimum), or the process whereby (1a) may result in an ironical comment on Joe's diplomatic skills, etc.. Such processes are arguably postcompositional, in the sense that they presuppose a grasp of sentence meaning, plus an understanding of the speaker's intentions, etc. We have no doubt that such processes exist (and, thus, that aspects of the Gricean picture are sound and effective). The question is whether SIs are phenomena of the latter postcompositional sort or are UG driven like, say, the principles of focus association sketched in (1).

1.1. Background

In his seminal work, Grice (1989) argues that the main source of pragmatic enrichment is a small set of maxims, summarized in schematic form in (2), that govern, as overridable defaults, conversational exchanges[2].

(2) **Quantity**
a. Make your contribution to the conversation as informative as is required
b. Do not make your contribution more informative than required

Quality
c. Do not say what you believe to be false
d. Do not say what you don't have adequate evidence for

Relation
e. Be relevant

Manner
f. Avoid obscurity and ambiguity
g. Be brief and orderly

In discussing the various ways in which these maxims may be used to enrich basic meanings, Grice considers the case of how *or* might strengthen its classical Boolean inclusive value to its exclusive construal. In what follows, we offer a reconstruction of the relevant steps of this enrichment process, as is commonly found in the literature (cf., e.g. Gamut (1991)). The basic idea is that, upon hearing something like (3a), a hearer

[1] The term 'postcompositional' has a possibly overly strong temporal connotation. But our point is not so much that all of pragmatic processes temporally follow the semantic one, as that they do not operate recursively on syntactic structures.

[2] Oswald Ducrot developed similar ideas independently (see e.g. Ducrot 1973), and also influenced early neo-Gricean works (e.g. Horn 1972, 1989, Fauconnier 1975a, 1975b).

considers the alternative in (3b) and subconsciously goes through the reasoning steps in (3i-vi)

(3) a. Joe or Bill will show up
b. Joe and Bill will show up
i. The speaker said (3a) and not (3b), which, presumably, would have been also relevant [relevance]
ii. (3b) entails (3a), hence is more informative[3]
iii. If the speaker believed that (3b), she would have said so [quantity]
iv. It is not the case that the speaker believes that (3b) holds
v. It is likely that the speaker has an opinion as to whether (3b) holds.
Therefore:
vi. It is likely that the speaker takes (3b) to be false.

This example illustrates how one might go from (3a) to (3vi) by using Grice's maxims and logic alone. The conclusion in (3vi) is close to the desired implicature but not quite. The conclusion we actually want to draw is that the speaker is positively trying to convey that Joe and Bill will not both come. Moreover, we need to be a bit more precise about the role of relevance throughout this reasoning, for that is a rather sticky point. We will do this in turn in the next three subsections.

1.2. SIs as exhaustifications.

To understand in what sense the conclusion in (3vi) should and could be strengthened, it is convenient to note that the reasoning in (3) can be viewed as a form of *exhaustification* of the assertion, i.e. tantamount to inserting a silent *only*. Using B_S as a short form for 'the speaker believes that', the assertion in (3a) would convey to the reader the information in (4a), while the alternative assertion in (3b) would convey (4b).

(4) a. B_S (show up(j) $\vee$ show up(b))
b. B_S (show up(j) $\wedge$ show up(b))

If you now imagine adding a silent *only* (henceforth, *O*) to (4a) (and evaluating it with respect to the alternative in (4b)), we get:

(5) O_{ALT} (B_S (show up(j) $\vee$ show up(b)))
= B_S (show up(j) $\vee$ show up(b)) $\wedge\neg$ B_S (show up(j) $\wedge$ show up(b))

The result in (5) is the same as (3iv) and entitles the hearer only to the weak conclusion in (3vi) (and for the time being, we might view this use of *O* as a compact way of expressing the reasoning in (3)). Now, the conclusion we would want instead is:

[3] Here and throughout we adopt the standard definition of the notion of strength, according to which p is stronger than q iff p asymmetrically entails q, though see our discussion in section 4.2.

(6) $B_S (O_{ALT} (\text{show up}(j) \vee \text{show up}(b)))$

$= B_S (\text{show up}(j) \vee \text{show up}(b) \wedge \neg (\text{show up}(j) \wedge \text{show up}(b)))$

The speaker, in other words, by uttering (3a), is taken to commit herself to the negation of (3b). The reading in (6) is tantamount to a kind of 'neg-raising' effect which allows us to go from something like *it is not the case that x believes that p* to *x believes that not p*. Sauerland (2005) calls this 'the epistemic step'. What is relevant in the present connection is that in the computation of SIs such a step does not follow from Gricean maxims and logic alone. It is something that needs to be stipulated. This seems to be a gap in the Gricean account of SIs (see for instance Soames 1982, Groenendijk and Stokhof 1984). And this problem interacts with another, even more serious one, having to do with seemingly innocent assumption that in uttering (3a), something like (3b) is likely to be relevant. Let us discuss it briefly.

1.3. Relevance

Let us grant that in uttering (3a), (3b) is also indeed relevant, whatever 'relevant' may mean. Now, a natural assumption is that the property of 'being relevant' is closed under negation, i.e. if a proposition ϕ is relevant, then $\neg \phi$ is relevant as well. To say that ϕ is relevant must be to say that it matters whether ϕ is true or false.[4] If this is so, the negation of (3b) will also be relevant. But then the set of alternatives changes. It minimally becomes:

(7) a. $B_S (\text{show up}(j) \vee \text{show up}(b))$

b. $B_S (\text{show up}(j) \wedge \text{show up}(b))$

c. $B_S (\neg(\text{show up}(j) \wedge \text{show up}(b)))$

Now, if we run the Gricean reasoning in (3) over this expanded set of alternatives or, equivalently, if we exhaustify the assertion along the lines discussed in (4) with respect to the alternatives in (7), here is what we get:

(8) $O_{ALT}(B_S (\text{show up}(j) \vee \text{show up}(b))) = B_S (\text{show up}(j) \vee \text{show up}(b))$

$\wedge\neg B_S (\text{show up}(j) \wedge \text{show up}(b))$

$\wedge\neg B_S (\neg(\text{show up}(j) \wedge \text{show up}(b)))$

This set of beliefs is consistent. It says that the speaker's only belief is that Joe or Bill will show up and that she has no opinion/evidence as to whether or not they both will

[4] An influential characterization of *relevance*, due to Carnap (1950), holds that a proposition ϕ is relevant relatively to a proposition ψ (ψ can be thought of as a proposition whose truth is under discussion) if the conditional probability of ψ relatively to ϕ is different from the (non-conditional) probability of ψ. The property of closure under negation follows from such a definition. One may also define relevance in terms of 'answerhood', as in Groenendijk and Stokhof (1984, 1990). Such a notion of relevance also enjoys the property of closure under negation.

show up. Notice that in this case, the epistemic step would lead to contradiction. I.e. we cannot 'neg-raise' our conclusions without imputing to the speaker contradictory beliefs.

Let us take stock. We have made a minimal and hard to avoid assumption on relevance: for any assertion *A*, if *B* is a potentially relevant alternative to *A*, then so is *not B*. This seemingly innocent move has the effect of blocking any potential SI. In particular, if the speaker utters *p or q* and *p and q* is relevant, then *not (p and q)* also cannot fail to be relevant. But then our Gricean reasoning in (3) yields that speaker must not know whether *p and q* holds.[5]

So, we see that on the one hand, by logic alone, we are not able to derive SIs in their full strengths from the Gricean maxims. And, if we are minimally explicit about relevance, we are able to derive no implicature at all (except 'ignorance' ones). Something seems to be going very wrong in our attempt to follow Grice's ideas. However, post Gricean scholars, and in particular Horn (1972, 1989), have addressed some of these problems and it is important to grasp the reach of such proposals. To this we now turn.

1.4. Scales.

Horn's important point is that if we want to make headway in understanding how SIs come about, then the set of relevant alternatives needs to be constrained. In the most typical cases, they will be *lexically* constrained by items of the same category whose entailments line them up in a scale of increasing informativeness. Examples of Horn's scales are the following:

(9) a. The positive quantifiers: *some*, *many*, *most*, *all*
b. The negative quantifiers: *not all*, *few*, *none*
b. Numerals: *one*, *two*, *three*, ….
d. Modals: *can*, *must*
e. Sentential connectives: *or*, *and*
f. Gradable adjectives: *warm*, *hot*, *boiling* / *chilly*, *cold*, *freezing*, etc.

[5] Strictly speaking, one also needs to assume that relevance is closed under conjunction (i.e. if *p* is relevant and *q* is relevant, then *p and q* is relevant). The reasoning then unfolds as follows.

i. *p or q* is relevant, because it is being asserted
ii. *p and q* is relevant by hypothesis
iii. *not (p and q)* is relevant because relevance is closed under negation
iv. *p or q and not (p and q)* is relevant because of (i), (iii) and the closure of relevance under conjunction.

Thus, from the fact that the speaker didn't utter (iv), which is both relevant (if (ii) is) and more informative than the assertion, we are forced to conclude that the speaker has no evidence that (iv) holds.

A precursor to this argument can be found in Kroch (1972); the point was worked out in detail by K. von Fintel and I. Heim in their 1997 class on pragmatics. For a consonant line of argumentation, cf. Davis (1998) .

These series are characterized by the fact that (under reasonable assumptions), the items on the right are stronger than the items on their left. For example, if all of the students did well, then most of them did and surely some of them did. Similarly for the other scales. Horn's proposal is that if you use *some*, other members of the scale may be activated and provide the alternatives against which the assertion is evaluated. Not all have to be activated; perhaps none of them will. But if they are activated, they must look like in (9). What is crucial about these scales is that one cannot mix elements with different monotonicity/polarity properties (see Fauconnier 1975b and Matsumoto 1995). Thus for example, one cannot have positive and negative quantifiers as part of the same scale. This is the way the problem considered in section 2.2. is circumvented.

Horn's suggestions can be extended to other seemingly more volatile/ephemeral scales. Consider the following example, modeled after Hirschberg (1985):

(10) A: Did John mail his check?
B: He wrote it.

This dialogue suggests that B's intention is to convey that John didn't mail the check. The 'scale' being considered here must be something like {write the check, mail the check}. What is crucial is that we do not consider mailing vs. not mailing, or mailing vs. stealing, for otherwise we would only derive ignorance implicatures.

The main moral is that the notion of 'relevance' to be used in implicature calculation is, yes, context dependent but constrained (by grammar, one would want to say) in at least two ways: through the lexicon (certain classes of words form lexical scales) and through a monotonicity constraint (all scales, even scales that are not lexically specified, such as those needed for (10), cannot simultaneously include upwards and downwards entailing elements).[6]

As mentioned, the goal of this paper is to challenge the neo-Gricean approach to SIs[7] based on the existence of embedded implicatures and to briefly introduce a few related considerations, all of which argue for an alternative *grammatical approach* to the

[6] We define here downward vs upward entailing and we also generalize the definition of entailment to non propositional types. Both definitions are standard.

(a) A function f is downward (respec. upward) entaling iff whenver $A \subseteq B$, $f(B) \subseteq f(A)$ (respect. $f(A) \subseteq f(B)$)

(b) Suppose that A and B are of the same non propositional type; suppose moreover, that $A(a_1)....(a_n)$ is a proposition; then $A \subseteq B \leftrightarrow \forall a_1, \ldots a_n [A(a_1)\ldots(a_n) \subseteq B(a_1)\ldots(a_n)]$.

We should add, furthermore, that there is a limitation to the monotonicity constrained discussed in the text. Such a limitation concerns the case of so called 'verum focus'. See Romero and Han (2004) and Guerzoni (2004) for relevant discussion. This limitation reinforces our view that the notion of relevance employed in this family of phenomena is grammatically governed. See also Katzir (2007).

[7] We use the term "neo-Gricean" to characterize theories that derive SIs from Grice's maxims of conversation, generally supplemented with the notion of *scale*, and view SIs as resulting from a reasoning process about speakers' intentions, such as Horn (1972, 1989), Fauconnier (1975a, 1975b), and Levinson (1983). See also, for a more philosophically oriented approach, Bach and Harnish (1979), and Bach (1994). Gazdar (1979) offered one of the first formally explicit neo-Gricean theories of SIs. More recently, several formal implementations of neo-Gricean ideas have been proposed (Spector 2003, 2006, 2007b; Sauerland 2004, van Rooij and Schulz 2004, 2006). Another approach, broadly inspired by Grice but which departs more radically from the original formulations, can be found within the tradition of Relevance Theory (Sperber and Wilson 1986, Carston 1988).

basic phenomena. We will begin in section 2 with an illustration of what a grammatical approach to SIs might look like, and we provide a a preliminary argument, based on a sub-case of the generalization mentioned in the introduction, namely that embedded implicatures are possible in downward entailing and non-monotonic contexts. Section 3 will provide a detailed and new argument for the existence of embedded implicatures in upward entailing contexts. Finally, section 4 will review other arguments that have been recently given for a grammatical approach to SIs.

2. Embedded Implicatures: a first crack in the Gricean picture

2.1. Exhaustification as a grammatical device

Does Grice's approach, emended as proposed by Horn, provide us with a framework in which SIs may be properly understood? Horn's move helps us get around the problem raised by the appeal to relevance; but the epistemic step remains unaccounted for: reasoning via the maxims about the speaker's intentions gets us at best from something like (11a) to (11b):

(11) a. John or Bill will show up
b. The speaker has no evidence that they both will show up

For the time being, let us simply assume/stipulate that the epistemic step can take place and that the speaker's intentions by uttering (11a) may be to convey something like (12a), which we represent as (12b):

(12) a. John or Bill and not both will show up
b. O_{ALT}(John or Bill will show up)

For concreteness, we may assume that if the alternatives are active (and hence the set ALT is non empty), such alternatives are obligatorily factored into meaning via O. Otherwise, if the alternatives are not active, the plain unenriched meaning is used, and no SI comes about. (This assumption will be motivated in section 4.1.)

The discovery that alternatives, when active, are constrained lexically and by monotonicity may also have a more radical effect on our perspective on SIs. So far, we have regarded our silent *only* as way to express, in compact form, Gricean reasoning of the type exemplified in (3). However, a different interpretation of O becomes available in light of the above considerations.[8] Notice that there is independent evidence that covert uses of *only* do exist. Consider for example the following dialogue:

[8] For the time being, we define the operator O, modeled on *only*, as follows: O_{ALT}(S) expresses the conjunction of S and of the negations of all the members of ALT that are not entailed by S. More formally:
$□O_{ALT}(S)□^{w} = 1$ iff $□S□^{w} = 1$ □ $\forall \phi \in ALT\ (\phi(w) = 1 \rightarrow □S□ \subseteq \phi)$
It follows that in principle, if S' is an alternative of S that does not entail S *but is also not entailed by it*, the negation of S' will be an implicature of S. See section 4.5.

(13) A: So, did you see the students?
B: I saw $[_{F}$ Joe and Sue]
where the constituent $[_{F}$ Joe and Sue] bears focal stress

In such a case, B's utterance will unambiguously convey that B saw *only* Joe and Sue. In spite of the fact that no overt *only* is actually present. What may be going on is that focus activates alternatives; active alternatives must be put to use and one option is via a covert occurrence of *only*. We may, then, assume that something very similar happens in the case of scalar alternatives. Contrast (13) with (14):

(14) A: So what happened when you went to school?
B: Well, I saw some of the students

Here the constituent *some of the students* is not specifically focused (presumably there is just a default focus on the VP as a whole). Yet, use of *some* has the capacity to activate the corresponding lexical scale; if this happens, it might be reasonably expected to prompt a covert exhaustification, by analogy with what happens in (13).

If this is on the right track, then exhaustification becomes more than just a way of expressing Gricean reasoning compactly. It becomes a grammatical device, just like, arguably, the covert use of *only* in (13), the main difference lying in how alternatives become active (via focus, or through a lexical route). What lends preliminary plausibility to this interpretation is the observation, stemming from Horn's work, that grammar seems to have a say in how alternatives are constrained. So, the overall result of exhaustification converges nicely with Grice's attempt. But it integrates it by filling in what otherwise appear to be missing pieces.

There is a further important point to make. Let us consider an example like (15).

(15) A: Who will come to the party?
B: I doubt that Joe or Sue will come.

Here no implicature comes about, even though the conjunctive statement *Joe and Sue will come to the party* must be relevant on the background of the question in A. The reason might be the following. The active alternative to B is *I doubt that Joe and Sue will come to the party*. Since B's utterance *entails* this alternative, the latter cannot be excluded and no implicature comes about. In our terms, use of covert *only* in cases like these is simply vacuous (as a straightforward computation will reveal to the reader). It is useful to compare (15) with:

(16) A: Who will come to the party?
B: I doubt that all of the students will
B-ALT: I doubt that (some of the) students will come to the party

In section 3.4.2., we will propose a modification of this operator. Quite often in this paper, we'll omit the subscript *ALT*, because we will make the simplifying assumption that the set ALT is predictable from the linguistic properties of the sentence to which O applies.

If B's answer is as indicated, its alternative would presumably be something like B-ALT.[9] So, exhaustifying B's utterance will bring about the negation of B-ALT, namely:

(17) It is not true that I doubt that (some of the) students will come to the party
= I believe that some of the students will come to the party.

This appears to be the right result: B's response to A does seem to implicate (17). This effect of scale reversal under negation, emphasized by several authors (cf., among others, Fauconnier 1975a, 1975b; Atlas and Levinson 1981, Horn 1989) is a welcome result. It generalizes Grices's insight to negative contexts. Our reformulation in terms of exhaustification adds nothing to it, other than a clear notation for cashing it in (but some independent motivation for exhaustification keeps, of course, resting on the grounds discussed above).

We are now in condition to properly address the main issue of the present paper. So far, we have been discussing implicatures that occur in unembedded contexts (like our original case in (3)). And even when we consider cases where implicature triggers (Horn Scale members, a.k.a. scalar items) occur in embedded position, as in (15) and (16), the relevant implicatures appear to be computed at the root level. This is in keeping with Grice's insight that implicatures arise by reasoning on speakers intention given a particular speech act (i.e., whole utterance).

The question we would like to ask now is whether SIs are *always* computed at the root level. If Grice is right, it should indeed be so. In this respect, the view we are developing that implicatures arise via something like a covert use of *only*, suggests that a different answer to the question might be right. For there is no a priori reason for thinking that covert uses of *only* are restricted to the root. If such an operator exists, it is unclear what should prevent its deployment at embedded scope sites. However, whether we are right or wrong on how SIs come about, the issue of whether they can systematically arise in embedded positions clearly deserves close inspection.[10]

Summing up, the Gricean view, emended a la Horn with grammatically based constraints on scales, and with a stipulation on the epistemic step, is able to derive basic SIs. However, such an approach seems to clearly predict that SIs are a root, postcompositional phenomenon. This prediction has a prima facie very nice result with 'scale reversal' cases such as those in (15)-(17). The question is whether it withstands further scrutiny.

[9] There are irrelevant issues having to do with the positive polarity character of *some* that may make B-ALT somewhat less than felicitous – whence the parentheses.

[10] Chierchia (2004), in a paper which started being circulated in 2000, partly building on ideas by Landman (1998), argued for the existence of embedded implicatures and concluded that SIs are derived by means of compositional rules which apply recursively to the constituents of a given sentence. Such a theory can be called a 'localist' theory of SIs. See also Récanati (2003) for a version of this position. Several works reacted to this proposal by formulating 'globalist', neo-Gricean accounts of some of Chierchia's empirical observations (see for instance Spector 2003, 2006, 2007b; Sauerland 2004; van Rooij and Schulz 2004, 2006; Russell 2006), i.e. accounts in which the SIs of a given sentence S result from the interaction of conversational maxims, the global meaning of S, and that of S's alternatives. Horn (2005) is a recent assessment of several aspects of this dispute, from the neo-Gricean standpoint. See also Geurts (to appear-a) for a nuanced defense of the globalist view. One of our goals here is to offer new arguments for the 'localist' view.

The rest of this section aims at giving a first argument for embedded implicatures, based on the existence of so-called ‘intrusive’ implicatures (see Levinson 2000), i.e. cases where a scalar item retains its ‘strengthened’ meaning under the scope of a downward entailing (DE) or non-monotonic (NM) operator. Embedded implicatures in upward entailing (UE) contexts will be the topic of the following section (section 3), in which we will provide a detailed new argument for their existence. Let us give some preliminary motivation for this split (which will be reviewed more analytically in the sections to follow). Implicatures embedded in DE or NM contexts, if present, would be relatively easy to detect by mere inspection of the truth conditions of the relevant examples. As we shall see shortly, embedding an implicature in a DE context leads to a statement that is weaker than the corresponding statement without the implicature. Thus, by checking that a sentence can be taken as true in a situation incompatible with the absence of the implicature, we can conclude that the implicature has to be there. Likewise, embedding an implicature in a non monotone context results in truth conditions that are independent of the corresponding statement without the implicature. Thus, again, by direct inspection of our intuitions we will be able to assess the presence or absence of the embedded implicature. By contrast, embedding an implicature in an upward entailing context leads to a proposition that is stronger than the literal meaning of the sentence. Such a statement, being stronger than the corresponding statement without the embedded implicature, will of course be compatible with it. Thus a case might be made that the relevant implicature is simply not there. This makes implicatures embedded in UE contexts harder to detect (and to motivate). Nonetheless, in section 3 we will present various tools that will allow us to see very clear consequences of the presence of such implicatures.

2.2. Implicatures embedded in DE and NM contexts

Sometimes scalar items receive an enriched interpretation under the scope of negation. Examples that seem to force such an enrichment are the following.

(18) a. Joe didn’t see Mary or Sue; he saw both.
b. It is not just that you *can* write a reply. You must.
c. I don’t expect that some students will do well, I expect that all students will.

The first example in (18a) receives a coherent interpretation only if the embedded *or* is interpreted exclusively. Similarly, the modal in (18b) has to be interpreted as *can though need not*, and the quantifier *some* in (18c) as *some thought not all*. For all the sentences in (18), in other words, it is as if the implicature gets embedded under the negative operator. In our notation, the LF of, e.g., (18a) could be represented as:

(19) not O_{ALT}(John saw Mary or Sue)

Examples of this sort have been widely discussed in the literature (especially in Horn 1985, Horn 1989); they seem to require either focal stress on the implicature trigger and/or strong contextual bias. Horn argues that cases of this sort constitute metalinguistic uses of negation, i.e. ways of expressing an objection not to a propositional content but to some aspect of a previous speech act. The range of possible speaker's objection can be very broad and concern even the choice of words or the phonology.

(20) a. This isn't car, it's a Ferrari
b. You don't want to go to [leisister] square, you want to go to [lester] square

In particular, with sentences like (18), the speaker objects to the choice of words of his interlocutor, presumably for the implicatures they might trigger.

While the phenomenon of metalinguistic negation might well be real (if poorly understood) there are other examples of DE contexts not involving negation that seem to require embedded implicatures. In what follows, we will consider several such cases, modelled mostly after Levinson (2000). To begin with, consider the contrast in (21).

(21) a. If you take salad or dessert, you'll be real full
b. If you take salad or dessert, you pay \$ 20; but if you take both there is a surcharge

In (21a) there is no implicature (*or* is construed inclusively); on the other hand, on the inclusive reading, (21b) would be contradictory. A coherent interpretation, which is clearly possible, requires an embedded implicature. [11] Let us go through the reasons why this is so. Suppose that in the context where (21b) is uttered, the alternative with *and* is active. Then, there may be in principle two sites at which the implicature is computed. Using our notation, they can be represented as follows:

(22) a. O_{ALT}(if you take salad or dessert, you pay \$ 20)
b. if O_{ALT}(you take salad or dessert), you pay \$ 20

If the option in (22a) is taken, the relevant alternative set would be:

(23) a. If you take salad or dessert, you pay \$ 20
b. If you take salad and dessert, you pay \$ 20
c. O_{ALT}(if you take salad or dessert, you pay \$ 20) =
if you take salad or dessert, you pay \$ 20
$\wedge \forall p\ p \in$ ALT [if you take salad or dessert, you pay \$ 20 $\not\subseteq p \rightarrow \neg p$]
= if you take salad or dessert, you pay \$ 20

Since the assertion (23a) is stronger than its alternative, the truth conditions of (22a) wind up being the same as those of (23a) (i.e. no implicature comes about – cf. the computation in (23c)). And as already noted, this reading is too strong to be compatible

[11] It might be objected that a non-monotonic analysis of conditionals (Stalnaker 1968, Lewis 1973) does not predict (21b) to be contradictory on the inclusive reading of disjunction. But according to such an analysis, (21b) should entail that the worlds most similar to the actual world in which the addressee takes salad or dessert but not both are more similar to the actual world than those in which he takes both. Yet clearly an utterance of (21b) is perfectly felicitous in a context where it is known that the addressee is more likely to have both salad and dessert than to have only one of them (because, for instance, this is always what he does). Hence we believe that the presence of an embedded implicature in (21b) is as clear as it is in (25c) below, which does not involve a conditional but a universally quantified statement.

with the continuation in (21b). So the LF in (22a) is unavailable. On the other hand, if the implicature is computed at the level of the antecedent, as in (22b), we get the equivalent of:

(24) If you take salad or dessert and not both, you pay $ 20

The truth conditions of (24) are weaker than those of (22a), and this makes them compatible with the continuation in (21b). Thus, the fact that sentences such as (21b) are acceptable seems to constitute prima facie evidence in favour of the possibility of embedding SIs.

This phenomenon seems to be quite general. Here are a few more examples involving the antecedents of conditionals, as well as further DE contexts like the left argument of the determiner *every*.

(25) a. If most of the students do well, I am happy; if all of them do well, I am even happier
b. If you can fire Joe, it is your call; but if you must, then there is no choice
c. Every professor who fails most of the students will receive no raise; every professor who fails all of the students will be fired.

It should be noted that these examples can be embedded even further.

(26) a. John is firmly convinced that if most of his students do well, he is going to be happy and that if all of them will do well, he'll be even happier.
b. Every candidate thought that presenting together his unpublished papers and his students evaluation was preferable to presenting the one or the other.

Without adding implicatures at a deeply embedded level, all of these examples would be contradictory.[12] For instance, in (26a) the implicature is embedded within the antecedent of a conditional, which is in turn embedded under an attitude verb.

A similar argument can be replicated for non monotonic contexts. Consider the following example:

(27) Exactly two students wrote a paper or ran an experiment.

It seems equally possible to interpret *or* in (27) inclusively or exclusively. It might depend on whether the requirement for the relevant class was an inclusive choice of a paper or experiment vs. say a presentation, or whether students were not allowed to write both a paper and run an experiment.

[12] One could imagine various semantic analyses of 'preferable' that would give (26b) a coherent interpretation even on an inclusive interpretation of disjunction. But the following observation casts doubt on such an attempt: a sentence of the form 'A is preferable to B' is generally felt as contradictory when A entails B, as the following illustrates.
(i) # Having a cat is preferable to having a pet.
So on an inclusive construal for disjunction, (26b) would be expected to be odd too.

Recall, now, that the truth conditions associated with the exclusive vs. inclusive construals of *or* under the scope of a non monotone quantifier as in (27) are independent of each other. For example, in a situation in which one student writes a paper and another writes a paper and also runs an experiment (and nobody else does either), the sentence is true on the inclusive construal of *or*, but false on the exclusive construal. On the other hand, in a situation in which one student only writes a paper , another only runs an experiment and other students do both falsifies the inclusive interpretation of *or* in (27); in such a scenario, (27) is only true on the (embedded) exclusive construal. The relevant reading can be forced via continuations of the following sort:

(28) Exactly two students wrote a paper or ran an experiment. The others either did both or made a class presentation.

For (28) to be coherent, the implicature must be computed under the scope of *exactly one*. Cases of this sort are pretty general and can be reproduced for all scalar items: sentence (29) below must be interpreted as ‘three students did most though not all of the exercises’

(29) Exactly three students did most of the exercises; the rest did them all

.

Taking stock, we have discussed a number of example sentences involving a variety of DE and NM contexts. Such sentences appear to have coherent interpretations that can only be derived if the implicature is added/computed at an embedded level (i.e. within the scope of a higher verb or operator). It should be noted that focal stress on the scalar item often helps the relevant interpretation. From our point of view, this is not surprising. The mechanism we have sketched for implicature calculation is, in essence, covert exhaustification, one of the phenomena triggered by focus. But it should also be noted that while focal stress is often helpful, it doesn’t appear to be always necessary. More generally, for the time being, we make no claim as to the frequency or marked status of embedded implicatures (but see our discussion in section 4.6 below). Our point is simply that they can and do naturally occur and that there are ways in which embedded implicatures can be systematically induced. This fact seems to be incompatible with the claim that SIs are a postcompositional semantic process, as the Gricean or Neo-Gricean view would have it. Of course, to establish our claim fully, we would like to be able to show that SIs can also be embedded in UE contexts. As readers might recall, the reason why we separated embedding under DE and NM operators from UE ones is merely one of convenience: the presence of implicatures embedded in DE or NM contexts can be established by mere inspection of the truth conditions. Embedding in UE contexts calls for more sophisticated methods.

3. A new argument for embedded implicatures in UE contexts: Hurford's constraint[13]

As has just been noted, embedded implicatures in UE contexts may yield readings that are strictly stronger than the readings that result from no implicature at all or from a 'globally derived" implicature. Furthermore, if embedded implicatures exist, we expect many sentences to be multiply ambiguous, depending on whether an implicature is computed in a given embedded position or not. In UE contexts, the various readings that are predicted are all stronger than the 'literal' reading. So, it may prove hard to detect the existence of a particular predicted reading as a separate reading (since if a certain reading R1 entails another reading R2, there can be no situation where R1 is true and R2 is false). In order to circumvent this difficulty, one would like to be able to construct cases where, for some reason, only one of the potentially available readings is licensed. In this section, we are going to do exactly this: we'll show that some sentences will *have* to contain a local exhaustivity operator in a UE context for a certain constraint (Hurford's constraint) to be met.

3.1. Hurford's constraint

Hurford (1974) points to the following generalization:

> Hurford's constraint (HC): A sentence that contains a disjunctive phrase of the form *S or S'* is infelicitous if *S* entails *S'* or *S'* entails *S*.[14]

This constraint is illustrated by the infelicity of the following sentences:

(30) a. # Mary saw a dog or an animal.
b. # Mary saw an animal or a dog.
c. # Every girl who saw an animal or a dog talked to Jack.

However the following example, which is felicitous, seems to be a counterexample to HC:

(31) Mary solved the first problem or the second problem or both problems

If *or* is interpreted inclusively, then clearly (31) violates HC, since 'Mary solved both problems' entails 'Mary solved the first problem or the second problem'. On the basis of such examples, Hurford reasoned that *or* has to be ambiguous, and that one of its readings is the exclusive reading. On an exclusive construal of the first disjunction in

[13] This section develops some arguments that were first presented in a seminar taught by D. Fox in 2004 (, especially subsection 3.4.2, cf. Fox 2004b), as well as by an MIT/Harvard seminar jointly taught by D. Fox and G. Chierchia in 2006 (especially subsections 3.1, 3.2 and 3.3) and in B. Spector's 2006 dissertation (especially subsections 3.4 and 3.5, cf. Spector 2006).

[14] 'entail' has to be understood under its generalized version, i.e. 'is included in', so as to be able to apply to pairs of non-propositional constituents.

(31), the sentence no longer violates HC. Gazdar (1979) noticed other cases where HC appears to be obviated, such as (32):

(32) Mary read some or all of the books

If *some* is just existential quantification and if *all* is universal quantification, then 'all of the books' entails 'some of the books', and (32) violates HC. By analogy with Hurford's reasoning about disjunction, one might conclude that *some* is ambiguous as well, and means *some but not all* on one of its readings. But Gazdar argued that multiplying lexical ambiguities in order to maintain HC misses an obvious generalization. Namely, the items that have to be analyzed as ambiguous in order to maintain HC are all scalar items, and the new meanings that are introduced correspond to the SIs that these items induce in simple contexts. Instead of assuming that these scalar items are ambiguous (which would, in effect, amount to a rejection of the whole neo-Gricean enterprise), Gazdar proposed to weaken Hurford's generalization in the following way:

> Gazdar's generalization: A sentence containing a disjunctive phrase *S or S'* is infelicitous if *S* entails *S'* or if *S'* entails *S*, unless *S'* contradicts the conjunction of *S* and the implicatures of *S.*

Let us look at a schematized version of (31):

(33) (A or B) or (A and B)

Let *S* be *A or B* and *S'* be *A and B*. The conjunction of S and its implicatures is *A or B but not both*, which contradicts *S'*. So the felicity of (33) is predicted by Gazdar's generalization.

Even though Gazdar did not spell-out an account for this generalization, one could interpret his observations as suggesting that violations of HC involve some kind of "implicature cancellation mechanism", in the sense that the second disjunct is used, so to speak, to cancel an implicature of the first disjunct (see Sharvit and Gajewski 2007). Instead of resorting to such a mechanism, we'll argue for the following:

- HC is correct as originally stated
- All the apparent violations of HC involves the presence of an implicature-computing operator within the first disjunct, ensuring that HC is met – hence the presence of a 'local implicature'.[15]

It is clear that something close to Gazdar's generalization follows from these two assumptions: suppose S2 entails S1; then 'S1 or S2' violates HC; yet 'O_{ALT}(S1) or S2" may happen to satisfy HC; this will be so if S1 together with its implicatures is no longer entailed by S2, which will be the case, in particular, if S2 contradicts S1 together with its

[15] In a way, we are extending Hurford's original account based on ambiguity to all scalar items. But, in contrast with Hurford, we do not assume any kind of *lexical* ambiguity. Rather, on our view, scalar items appear to be ambiguous because of the optional presence of an embedded implicature-computing operator.

implicatures. For instance, a sentence of the form *A or B or both A and B* has to have the following underlying structure for it to satisfy HC: *[O_{ALT}(A or B)] or [both A and B]*.[16]

This analysis turns out to make very precise predictions in a number of cases – predictions that do not fall out from Gazdar's proposal. In the next subsections, we'll spell out these predictions and corroborate them in various ways. Since they crucially depend on the assumption that an embedded implicature-computing operator is obligatorily present in all the relevant cases, these predictions provide important evidence, in our view, for the claim that SIs can be computed in embedded positions.

3.2. Forcing embedded implicatures.

Gazdar's generalization, as such, does not make any particular prediction regarding the *reading* that obtains when there is an apparent violation of HC. But consider now the following sentence (in a context where it has been asked which problems Peter solved within a certain set of problems):

(34) Peter either solved both the first and the second problem or all of the problems.

In the absence of an exhaustivity operator, (34) would violate HC, since solving all of the problems entails solving the first one and the second one. And (34) would then be equivalent to (35):

(35) Peter solved the first problem and the second problem.

Therefore, we predict that an exhaustivity operator has to be present, with the effect that (34)'s logical form is the following:

(36) O_{ALT}(Peter solved the first problem and the second problem) or he solved all of the problems

Recall that the meaning of our operator is supposed to be – at least as a first approximation – the same as that of *only*.[17] If we are right, (34) should therefore be understood as equivalent to the following:

(37) a. Peter only solved the first problem and the second problem, or he solved all of the problems
b. Either Peter solved the first problem and the second problem and no other problem, or he solved all the problems

[16] In this paper, we do not account for the following asymmetry, pointed out by Singh (2006):
(a) Mary saw Peter or Sue or both Peter and Sue
(b) #Mary saw both Peter and Sue or Peter or Sue
See Singh (2006) for an interesting proposal.

[17] Recall that $O_{ALT}(\phi)$ expresses the conjunction of ϕ and the negations of all of ϕ's alternatives that are not entailed by ϕ. In the case at hand, we assume that the alternatives of "Peter solved the first problem and the second problem" consist of all the propositions of the form *Peter solved X*, where X is one of the problems or a plurality made up of some of the problems. So the alternatives, in this case, are not solely defined in terms of scales.

It turns out that this is indeed the only possible reading of (34). In other words, (34) is clearly judged *false* in a situation in which Peter solved, say, the first problem, the second problem and also the third problem (out of a set of more than three problems). It is hard to see how an analysis based on the notion of "implicature cancellation" could account for the particular interpretation that such a sentence triggers. Under such a theory, the presence of on implicature cancellation device is not expected to yield any new implicatures.[18] So example (34) seems to be a clear case of an embedded implicature.

It is worth noticing that the reading we observe is maintained when the whole sentence is itself embedded, say, in a DE-context:

(38) Whoever solved the first and the second problems or solved all of the problems will pass.

This sentence is understood as equivalent to:

(39) Whoever either solved only the first and the second problems or all of the problems will pass.

In short, our argument can be summed up as follows: if HC is correct, as originally formulated, and if the implicature-computing operator O can occur in embedded positions, then in some cases, the only way to satisfy HC is to insert O locally, and this gives rise to particular readings which turn out to be the only possible reading.

3.3. Forcing even more embedded implicatures

So far, we have discussed sentences of the form *S or S'* such that *S' entails S* but *S'* is incompatible with $O_{ALT}(S)$, to the effect that $O_{ALT}(S)$ *or S'* does not violate HC. Such sentences *force* the insertion of O_{ALT} within the scope of the main disjunction, thus making a case for embedded implicatures in UE contexts. We would like to show that there may also be cases where the only way to satisfy HC is to insert an exhaustivity operator in an even more embedded position, namely *within* a subconstituent of the first disjunct.

Consider the following sentence:

(40) Every student solved some of the problems

There have been discussions in the recent literature as to whether *some* in a sentence like (40) can be read as *some but not all* (e.g. Chierchia 2004, Spector 2006, Geurts, to appear

[18] Note indeed that no implicature gets truly cancelled; in the absence of the second disjunct, the first disjunct (*Peter solved the first problem and the second problem*) implicates that Peter didn't solve any other problem; once the second disjunct is added, the implicature doesn't really disappear; rather, it is integrated into the meaning of the first disjunct. If no exhaustivity operator applied to the first disjunct, there would be no simple way to derive the reading we get by applying an exhaustivity operator to the whole sentence:

(30) O_{ALT}(Jack solved the first problem and the second problem or all of the problems)
= Jack solved the first problem and the second problem and no other problem

-a). On the grammatical view we are considering, the possibility of such an interpretation is expected, since inserting O within the scope of *every student* would generate exactly this reading.[19] Using the same technique as before, we are going to show that we can force *some* in (40) to be read as *some but not all* by embedding (40) itself in a bigger structure.

Before going through the argument, let us make some preliminary points. Suppose (40) competes only with (41)

(41) Every student solved all of the problems

Then $O_{ALT}((40))$ is equivalent to (40) $\wedge \neg$(41), i.e:

(42) Every student solved some of the problems and at least one student did not solve them all

This is clearly a natural interpretation of (40).

Inserting O below *every student* gives rise to a stronger reading:

(43) a. (Every student)$_x$ (O_{ALT}(x solved some of the problems)
b. (Every student)$_x$ (x solved some of the problems and $\neg$(x solved all of the problems))
c. Every student read some of the problems but not all of them

Let us now construct a sentence in which (40) occurs as a constituent and is forced to correspond to the logical form in (43) for the whole sentence to comply with HC. We claim that the sentence given in (44) achieves exactly that:[20]

(44) It is either the case that Every student solved some of the problems, or that Jack solved all of them and all the other students solved only some of them.

Our empirical claim is that (44)'s only reading is one in which *some of the problems* in the first disjunct is interpreted under its strengthened meaning, i.e. *some but not all*. In other words, we claim that (44) is equivalent to (45), so that it is judged *false* if a student other than Jack solved all of the problems:

(45) Every student solved some of the problems, and no student except maybe Jack solved all of the problems.

[19] On the globalist view, this interpretation should not necessarily be ruled out: some globalist theories (van Rooij & Schulz 2004, 2006, Spector 2003, 2006, 2007b) are able to generate this 'strengthened' meaning, provided the following also count as alternatives of (40): "some students solved some of the problems", "some students solved all of the problems". Fox (2007) proposes a constraint on alternatives that rules out such a large set of alternatives for (40) (cf. his footnote 35). For the sake of this discussion, we are going to assume that the *some but not all* reading under a universal quantifier can only be derived as a local implicature.

[20] Some speakers might feel that in (44) *some* has to be stressed.

This equivalence follows directly from the combination of HC and the possibility of inserting O_{ALT} locally. Let us show why.

First consider what happens if (44) contains no exhaustivity operator at all: then clearly HC is violated since the second disjunct entails the first one. What if we apply O_{ALT} to the first disjunct, as in the previous examples?

(46) O_{ALT} (Every student solved some$_F$ of the problems), or Jack solved all of them and every other student solved only some of them

This logical form is equivalent to the following sentence:

(47) Every student solved some of the problems and not all of the students solved them all, or Jack solved all of them and every other student solved only some of them

It turns out that this still violates HC: the second disjunct does in fact entail that not all of the students solved all of the problems, hence entails the first disjunct. So the only possible analysis is the following:

(48) Every student$_x$ (O_{ALT}(x solved some$_F$ of the problems) or Jack solved all of them and all the other students solved only some of them.

(48) yields exactly the reading that we in fact observe, which, notice, is neither equivalent to the 'literal' reading of the sentence (the reading that we would see if there were no operator at all) nor to the 'globally strengthened' meaning of the sentence (the reading that we would see if there were just one operator scoping over the entire sentence).[21] Not only have we constructed an example where an embedded implicature clearly arises[22], but we have also provided a precise account of the manner by which this reading comes about. This account, if successful, provides support to our two assumptions, i.e. that a covert exhaustivity operator can be inserted locally and that HC is correct as originally formulated.

3.4. HC and recursive exhaustification: when a locally inserted operator gives rise to new 'global' implicatures

In this section, we are going to show that even in cases where the obligatory presence of an embedded implicature-computing operator does not have any direct effect on the

[21] This last claim cannot be rigorously defended without being quite explicit about the alternatives of such a complex sentence, and we do not provide a complete argument in this paper. Intuitively, what happens is that (44) (in the absence of a covert exhaustivity operator) and (40) are equivalent and, given certain reasonable assumptions about how alternatives are computed, they remain equivalent even after global exhaustification, i.e. O_{ALT}((44)) = O_{ALT}((40)) (=*Every student solved some of the problems, and not all students solved all of the problems*).

[22] Note that the claim that (44)'s interpretation involves an embedded implicature does not in fact depend on our assumption that the *some but not all* reading of *some* requires the presence of the operator within the scope of the universal quantifier. As mentioned in fn.19, it may be possible to derive this reading by applying the operator to the first disjunct as a whole. What is crucial here is that there must be an exhaustivity operator **within** the first disjunct, hence in an embedded position.

literal truth-conditions of a given sentence, it nevertheless has observable effects that can be detected by looking at the *implicatures* (or lack thereof) that the sentence in question triggers. In our terms, the presence of the embedded implicature-computing operator turns out to have a truth-conditional effect when the sentence is itself embedded under another implicature-computing operator. First, we'll look at the interpretation of disjunctions of the form 'A or B or both' in the scope of necessity modals (3.4.1).[23] Then we'll offer an account of the 'cancellation effect' triggered by *or both* in non-embedded contexts (3.4.2).

3.4.1.*Or both* in the scope of necessity modals

Consider the following two sentences:

(49) We are required to either read *Ulysses* or *Madame Bovary*
(50) We are required to either read *Ulysses* or *Madame Bovary* or both

What is clear about these sentences is that they both implicate that we are not required to read both *Ulysses* and *Madame Bovary*.[24] Otherwise, at first sight, they do not seem to trigger different implicatures. But upon reflection, they, in fact, do. Suppose that we are actually required to read *Ulysses* or *Madame Bovary* and that we are not allowed to read both of them. Then (49) would be an appropriate statement, while (50) would not. (49), on its most natural reading, is silent as to whether or not we are allowed to read both novels. But (50) strongly suggests that we are allowed to read both novels. So (50) seems to trigger an implicature that (49) does not.

This is further confirmed by looking at the following dialogues:[25]

(51) A: We are required to either read *Ulysses* or *Madame Bovary*
B: No! we have to read both
(52) A: We are required to either read *Ulysses* or *Madame Bovary*
B: ## No! We are not allowed to read both
(53) A: We are required to either read *Ulysses* or *Madame Bovary* or both
B: No! We are not allowed to read both

(51)B shows that in a denial context, as we have noticed in section 2, negation can target an implicature: B, in (51), is objecting to an implicature of the previous sentence, namely, that we are not required to read both novels. (52)B is clearly deviant. This shows that

[23] Section 3,4,1 is of a higher order of complexity than the rest of the arguments developed here. We present it because we find it particularly strong. However, its understanding is not crucial for our main point.

[24] Interestingly, the presence of *or both* does not cancel the implicature that is normally triggered in the absence of *or both*. This, in our view, is unexpected if *or both* is viewed as an implicature-cancelling device, a view that might be attributed to Gazdar (1979), as we mentioned above.

[25] The presence of *either* is not crucial to these judgments; its function here is simply to rule out the wide scope interpretation of disjunction, which is irrelevant here (see Larson 1985). In particular, note that *either…or* does not generally force an exclusive reading, as illustrated by the fact that the following sentence is perfectly consistent: 'We are required to either read *Ulysses* or *Madame Bovary*, and we may read both'

(52)A does *not* implicate that we are allowed to read both novels; indeed, if (52)A did trigger this implicature, then B's objection would be perfectly felicitous, since it would count as an objection to an implicature of A's utterance. What is important in the current context is that (52) clearly contrasts with (53). B's objection in (53) is completely natural, and hence confirms our claim that (53)A does implicate that we are allowed to read both novels.

How are these facts to be explained? How come (50) has an implicature that (49) doesn't have? Note that (modulo the presence of matrix *O*) (49) and (50) have the same truth-conditions. Yet they trigger different implicatures. We are going to show that this phenomenon is in fact entirely expected from our perspective. The implicatures associated with (49) and (50) are, in fact, instances of the following generalization:

(54) A sentence of the form $\Box(A \text{ or } B)$ triggers the following implicatures:[26]

$\neg \Box A, \neg \Box B$

To illustrate this generalization let us begin with (49), repeated here as (55), which implicates that we have a choice as to how to satisfy our obligations.

(55) We are required to either read *Ulysses* or *Madame Bovary*

The reading of (55), 'pragmatically strengthened' on the basis of the generalization in (54), is:

(56) We are required to either read *Ulysses* or *Madame Bovary*, we are allowed to read *Ulysses* without reading *Madame Bovary*, we are allowed to read *Madame Bovary* without reading *Ulysses*.

This happens to be exactly equivalent to the conjunction of (55) and the following propositions:[27]

(57) We are not required to read *Madame Bovary*
(58) We are not required to read *Ulysses*

[26] $\Box$ stands for any modal operator with universal force (*we are required to* …, *Necessarily* … , *Jack demanded that…, etc.*)

[27] Proof:

i) Suppose (56) is true. Then a) we are required to read one of the two novels and b) since we are allowed to read either one without reading the other one, we are not required to read *Ulysses* and we are not required to read *Madame Bovary*

ii) In the other direction. Suppose that (55), (57) and (58) are true. By (55), in all the permissible worlds we read one of the two novels; by (58) in one permissible world w_1 we don't read *Ulysses*; but then in w_1 we read *Madame Bovary* (by (55)), hence we read *MB* without reading *U*, and therefore we are allowed to read *MB* without reading *U*; Symmetrically, from (57) it follows that we are allowed to read *U* without reading *MB*. It follows that we are not required to read both.

Now let's see the consequences of this generalization for the case of (50) ("We are required to read *Ulysses* or *Madame Bovary* or both"), schematized as follows:

(59) $\Box$ $[O_{ALT}(A \text{ or } B) \text{ or } (A \text{ and } B)]$

(59) is predicted to implicate the following:

(60) $\neg \Box$ $(O_{ALT}(A \text{ or } B))$
(61) $\neg \Box$ $(A \text{ and } B))$

i.e. :

(62) We are not required to read *Ulysses* or *Madame Bovary* and not both
(63) We are not required to read both *Ulysses* and *Madame Bovary*

The end-result is the following proposition, which indeed corresponds to the most natural reading of (50):

(64) We are required to read *Ulysses* or *Madame Bovary*, we are not required to read only one of the two novels, we are not required to read both novels.

From (64) it follows that we are allowed to read both novels, which was the desired result. So far we have shown that given the generalization in (54), the observed interpretation of (50) follows directly from the assumption that O is present in the first disjunct. Of course, it is also desirable to understand why this generalization holds, in order to give a complete account. It turns out that the exhaustivity operator as we have defined it so far (adding the negations of all non-weaker alternatives) can derive all the inferences that we observed, provided we now assume that the scalar alternatives of a disjunctive phrase *X or Y* include not only the phrase *X and Y* but also each disjunct *X* and *Y* independently. If so, then (55) has the following alternatives (using now the sign '$\Box$' to abbreviate 'we are required to', 'U' to abbreviate 'read *Ulysses*', and 'MB' to stand for 'read *Madame Bovary*'):[28]

(65) ALT((55))= $\{\Box(U \text{ or } MB), \Box U, \Box MB, \Box(U \text{ and } MB)\}$

[28] Following standard assumptions, the scalar alternatives of a complex expression are derived by possibly multiple substitutions of scalar terms with one of their scale-mates. For any sentence ϕ containing scalar terms $<t_1,\ldots,t_2>$ (noted $\phi(t_1, \ldots, t_n)$), the alternatives of ϕ are all the sentences of the form $\phi(t'_1,\ldots, t'_n)$, where t'_j belongs to the scale of t_j for any j. What we need here is that the alternatives of any sentence ϕ containing *X or Y* (noted ϕ*(X or Y))* include $\{\phi(X \text{ or } Y), \phi(X), \phi(Y), \phi(X \text{ and } Y)\}$. Technically, one possible way of deriving this alternative set is to assume, as in Sauerland (2004), that the scale of *or* is the following partially ordered set : <*or, L, R, and*> where *L* is the binary connective such that *X L Y* is equivalent to *X*, and *R* is the one such that *X R Y* is equivalent to *Y*. For alternatives, see Spector (2006) and Katzir (2007).

To get the ‘strengthened meaning’ of (55), we simply have to add the negation of each of its alternatives that is not weaker than it. In this case, this results in the following, from which the generalization in (54) follows:

(66) $O_{ALT}((55)) = \Box(\text{U or MB}) \wedge \neg\Box\text{U} \wedge \neg\Box\text{MB}$

Let us now go back to our original example:

(67) a. We are required to either read *Ulysses* or *Madame Bovary,* or both
b. $\Box(O_{ALT}(\text{U or MB}) \text{ or } (\text{U and MB}))$

By assumption, the logical form of (67)a must be (67)b (i.e. the exhaustivity operator must have been introduced for HC to be satisfied). Given our new assumptions about the alternatives of a disjunctive phrase, (67)’s alternatives include, among others, ‘$\Box(O_{ALT}(\text{U or MB})$’ (i.e. “we are required to read either *Ulysses* or *Madame Bovary* and we are forbidden to read both”) and ‘$\Box(\text{U and MB})$’.

To prevent the discussion from getting too long, let us make the simplifying assumption that these are the only alternatives. As these two alternatives are not logically weaker than (67) (which is actually equivalent to “we are required to read *Ulysses* or *Madame Bovary*’), the strengthened meaning of (67), i.e. what results from an application of the operator O to the whole sentence, has to entail the negation of these two alternatives, which gives rise to a reading paraphrased in (68) (in three different ways). This is exactly what we wanted to derive.

(68) a. $\Box(\text{U or MB})) \wedge \neg\Box(O_{ALT}(\text{U or MB})) \wedge \neg\Box(\text{U and MB})$
b. We are required to either read *Ulysses* or *Madame Bovary,* and we are not required to read only one of them, and we are not required to read both
c. We are required to read *Ulysses* or *Madame Bovary* and we are allowed to read both of them and we are allowed to read only one of them

Let us sum up what has been shown: the obligatory presence of an exhaustivity operator applying to the first disjunct in sentences like (50), together with the assumption that each member of a disjunctive phrase contributes to the alternatives of the disjunctive phrase, immediately predicts that (50), though equivalent to (49), has more alternatives. The existence of these additional alternatives predicts, in turn, that the *strengthened* meaning of (50) is different from that of (49), and the precise prediction that is made seems to be corroborated.[29]

[29] Recall, however, that we made the simplifying assumption that (59) had only two alternatives. When we add the alternatives that are induced by the more embedded disjunctions and conjunctions, we end up with a much bigger set (even after having eliminating some alternatives that are equivalent to other ones): $\{\Box\,(\text{U or MB}), \Box O_{ALT}(\text{U or MB}), \Box O_{ALT}\text{U}, \Box O_{ALT}\text{MB}, \Box(\text{U and MB}), \Box\text{U}, \Box\text{MB}\}$. As a result, the following inferences are also predicted: we are not required to read both novels, we are not required to read *Madame Bovary*, we are not required to read *Ulysses*. These are good results as well. Note however that we could get a different result if we considered even more alternatives, namely, those that are triggered by the scale <*allowed*, *required*>. Fox (2007) proposes a constraint on the computation of alternatives that rules out these additional alternatives.

One crucial assumption for this account was that the alternatives of a disjunctive phrase include each of the disjuncts separately. This move, which is independently motivated[30], gives rise to a certain problem: once this assumption is made, the exhaustivity operator as we have defined it so far[31] leads to the contradictory proposition when applied to a simple, non-embedded disjunctive sentence.[32] We'll address this problem in the next sub-section by the introduction of a new, more realistic, exhaustivity operator. But what is important so far is that this more realistic operator allows us to maintain the account that has just been presented: it turns out that for all the cases that we have considered in this subsection, the operator O, as we have defined it initially, yields the same result as the more realistic operators that have been proposed in the recent literature.[33]

3.4.2 Cancellation

In this section, we show that the assumptions that we have developed in the previous sections are also able to predict the most well-known observation about the behavior of *or both*. Namely, in non-embedded contexts, 'or both' *removes* the exclusive reading of disjunction ((69) below, schematized in (70)), i.e. seems to 'cancel' an implicature.

(69) Peter or Jack or both came
(70) O_{ALT}(p or j) or (p and j)

In our terms, this is again a case where the presence of a local exhaustivity operator has no direct detectable truth-conditional effect (at least as far as the basic sentence is concerned), since (70) <*(p or_{excl} q) or_{incl} (p and q)*> is equivalent to *(p or_{incl} q)*. As before, though, the presence of an embedded exhaustivity operator creates more alternatives. We will now show that these new alternatives affect the interpretation that result from embedding this basic sentence under (yet another) exhaustive operator. Specifically, applying *O* to (70) is a vacuous move, which accounts for the 'cancellation effect' triggered by *or both*. In order to show how this prediction comes about, it will be necessary to use a more realistic exhaustivity operator, one that does not yield contradictions when it applies to simple disjunctive sentences (see fn. 32).

Before doing so, though, let us notice that once it is assumed that the presence of *or both* entails the presence of an implicature-computing operator applying to the first disjunct, the 'cancellation' effect of *or both* happens to be an instance of the following generalization (which is due to Gazdar 1979):

[30] This assumption has been shown to be able to solve a much discussed puzzle regarding sentences in which a scalar term occurs under the scope of a disjunction. The puzzle was first discussed in Chierchia 2004. Sauerland 2004 offered a solution based on a modified scale for disjunction. Sauerland's ideas will be presented in the next subsection.

[31] See fn. 8.

[32] Here is why: if the alternatives of 'p or q' are {p or q, p, q, p and q}, then negating all the non-weaker alternatives amounts to adding $\neg$p, $\neg$q, $\neg$(p and q). But '(p or q) $\wedge \neg p \wedge \neg q \wedge \neg(p \wedge q)$' is contradictory.

[33] This is so thanks to the following fact: let O be the exhaustivity operator as we have defined it. Let O' be one of the operators defined in the recent literature (Spector 2003, 2006, 2007b; van Rooij and Schulz 2004, 2006; Fox 2007). Then, for any given sentence ϕ, **if $O_{ALT}(\phi)$ is not contradictory, then $O_{ALT}(\phi) = O'_{ALT}(\phi)$.**

(71) A sentence of the form ϕ *or* ψ triggers the inference that its author does not know whether ϕ is true and does not know whether ψ is true.

Applied to (70), this results in the following:

The author of (70) does not know whether *p or*$_{excl}$ *q* is true and does not know either whether *p and q* is true.[34] Now this conclusion is obviously incompatible with the derivation of an 'exclusive' implicature (i.e. an inference that the speaker believes that 'p and q' is false). Hence the 'cancellation' effect of *or both* follows.

In order to explain why the generalization in (71) is in fact correct, and what role the presence of the embedded implicature-computing operator plays in deriving the right result, it is necessary to move to a more sophisticated theory of implicature computation.

Let us first have a look at a simple disjunctive sentence, adding the crucial assumption that each disjunct is an alternative of the initial sentence:

(72) Peter or Jack came
(73) p or j

The alternatives of (p or j) are now assumed to be {p or j, p, j, p and j}. As explained in fn. 32, applying O_{ALT} to (73) relatively to this set of alternatives results in the contradictory proposition (because *p* and *j* are both strictly stronger than *p or j*, and we end up with $O_{ALT}(p\ or\ j) = (p \vee j) \wedge \neg p \wedge \neg j$). Plainly, some modification is needed.

In order to understand how the operator should be modified, it is useful to take a small digression and see how this problem can be dealt with within a neo-Gricean setting. So let us go back to our informal presentation of the neo-Gricean reasoning and run it on this case (we are now presenting the analysis found in Sauerland 2004 – related proposals can be found in Spector 2003, 2006, 2007b and van Rooij & Schulz 2004, 2006):

a- the speaker has said 'p or j', so she believes that p or j
b. It follows from the maxim of quantity that that the speaker does not believe more than this relative to the set of alternatives: in other words, she *only* believes 'p or j', i.e. she does not have the belief that p, she does not have the belief that j and she does not have the belief that p and j (this last statement is actually entailed by the two first)

At this point we have derived inferences of the form 'the speaker does not have the belief that S', which Sauerland terms *primary implicatures* (*secondary implicatures* are inferences of the form 'the speaker believes that not-S'). Now comes a crucial observation, based on the assumption that the speaker is logically coherent: since the speaker believes that 'p or j' is true and does not have the belief that p is true (primary

[34] Note that in the absence of the operator O, applying the generalization in (71) to (*p or*$_{incl}$ *q) or*$_{incl}$ *(p and q)* (i.e. to an LF that does not contain O) results in a contradiction, if the speaker is taken to believe what he says; for the speaker, on the one hand, would on the one hand have to believe that (*p or*$_{incl}$ *q) or*$_{incl}$ *(p and q)* is true. i.e. that *p or*$_{incl}$ *q* is true, and, on the other hand, would have to be undecided as to whether or not *(p or q)* is true.

implicature), she cannot have the belief that j is false. Indeed, if she believed 'not j', then given that she believes 'p or j', she would have to believe p. Symmetrically, the speaker does not have the belief that p is false. The end result of this reasoning is that the speaker has no opinion as to whether p is true or false and as to whether j is true or false. It is a fact that these 'ignorance inferences', which Gazdar 1979 called 'clausal implicatures', are indeed triggered by disjunctive statements in normal contexts.

As was explained in section 1.2., within the neo-Gricean framework, 'secondary implicatures' of the form "the speaker believes that not S", where S is an alternative of the sentence uttered, require an additional step, which allows us to move from "The speaker does not believe that p" to "The speaker believes that not-p". But it is now clear that sometimes this move will contradict previously established implicatures. For instance, in the above case, moving from "the speaker does not have the belief that p" to "the speaker believes that not-p" actually conflicts with the already established conclusion that "the speaker does not have the belief that not-p". Sauerland's proposal is that the move from *the speaker does not believe that p* to *the speaker believes that non-p* occurs if and only if this does not contradict the primary implicatures that have been derived in the first step. In the case of *p or j*, this move can apply to the alternative *p and j*, but not to *p* or to *j* taken in isolation. We end up with an exclusive reading for disjunction, together with two "ignorance inferences", according to which the speaker does not know whether p is true and does not know whether j is true.

Can we achieve the same results in a theory in which SIs are generated by an implicature-computing operator, rather than by purely pragmatic reasoning? What is needed is a more sophisticated definition of the operator O, one that ensures that the application of O to a given sentence S returns the proposition that *would* have resulted from the application of Sauerland's purely pragmatic procedure[35]. Various and related proposals in the recent literature provide us with such a definition.[36] We are not going to present these fairly technical proposals, and we will simply assume that the proposition that O returns when applied to a sentence S is what would result from applying Sauerland's procedure to S.

Let us therefore apply Sauerland's procedure to (69) - (70), repeated as (74)- (75)

(74) Peter or Jack came, or both came
(75) O_{ALT}(p or j) or (p and j)

[35] Though O does not derive primary implicatures. Those would probably be pragmatic even in a theory that incorporates O. See section 4.4. for cases where predictions can diverge due to this difference between Sauerland's procedure and a theory derives many of the same results by a particular meaning for O.

[36] See Spector (2003, 2006, 2007b), van Rooij & Schulz (2004, 2006), Fox (2007). Spector's and van Rooij & Schulz's proposals are heirs to Groenendjik & Stokchof's (1984) exhaustivity operator. For the time being, it is enough to assume that a non-weaker alternative gets negated if and only if negating it is consistent with negating any other non-weaker alternative. In the case of 'p or j', the only such "innocently excludable" alternative is actually 'p and j'. An actual formalization in these terms is given by Fox (2007). Fox's definition of O is the following:

$O_{ALT}(\phi) = \phi \wedge \neg a_1 \wedge \neg a_2 \wedge \ldots . \wedge \neg a_n$

Where $\{a_1, \ldots, a_n\}$ is the set of all *innocently ϕ– excludable members of ALT* (where *ALT* is the set consisting of the alternatives of ϕ)

Def: An alternative of ϕ is innocently ϕ -excludable if its negation belongs to all maximal-consistent sets that include only ϕ and negations of an alternative of ϕ

(75)'s alternatives include, among others, the following two sentences: *O_{ALT}(p or j)* (which is identical to *p or j but not both*) and *p and j*. Both these sentences are strictly stronger than (75) (which is equivalent to *Peter or$_{incl}$ Jack came*). We therefore derive the following primary implicatures: *the speaker does not have the belief that O_{ALT}(p or j)*, and *the speaker does not have the belief that p and j*. Can we "lower" these negations without generating a contradiction? We would end up with the following inferences::

- p or j
- $\neg$ (p and j)
- $\neg$((p or$_{excl}$ j) = ($\neg$p and $\neg$ j) or (p a j)

But these three statements, taken together, are contradictory (if 'p or j' is true, then either '(p or j) and $\neg$(p and j)' or 'p and j' is true). It follows that none of these two alternatives is 'innocently excludable' (following Fox's 2007 terminology). So the second step will actually be vacuous, and no secondary implicature will be derived. The reader will easily check that this remains true even when we take into considerations the other alternatives (namely: {p, j, p or j, O_{ALT}(p), O_{ALT}(j), O_{ALT}(p and j)}. So we have shown the following:

(76) O_{ALT}((75)) = (75)

Note that in the absence of an exhaustivity operator applying to the first disjunct, we would in fact derive *$\neg$(p and j)* as a secondary implicature: for the alternatives of *p or q or both* would then be {p, q, p or q, p and q}[37], i.e. exactly the same as the ones for *p or q*.

3.5. Conclusions

In this section, we have offered a theory of the interaction of SIs and Hurford's constraint. The hypothesis that embedded implicatures can be generated freely in DE contexts as well as in non-DE contexts, seems to us to be needed in order to predict a wide array of facts pertaining to the interpretation of sentences which *apparently* violate HC but in fact do not, once the possibility of embedding an implicature-computing operator is acknowledged.

We used HC to force the presence of an exhaustivity operator in an embedded position, and were able to show that in some cases, very specific readings were predicted which turned out to be the only possible readings. In other cases, the obligatory presence of a local exhaustivity operator did not alter the truth-conditions *obligatoriy*, but did so once the whole sentence was itself embedded under another exhaustivity operator: the presence of the embedded operator sometimes yields additional implicatures, or, in other cases, prevents some implicatures from arising. From this perspective, the phenomenon of 'implicature cancellation' by means of (apparently) redundant disjuncts like *or both* or *or all* turned out to be a sub-case of a wider observation: the presence of an embedded

[37] $(p \vee q) \wedge (p \wedge q)$ can be eliminated, as it is equivalent to $(p \wedge q)$.

operator, even when it does not have direct truth-conditional effects, can affect the sentence's *alternatives* with possible consequences for the resulting interpretation once an additional alternative-sensitive operator is introduced. The corroboration of these complex predictions supports the main premise, i.e. that an implicature-computing operator can be inserted locally.

4. Further Cracks in the Gricean picture

In the previous sections we have seen various reasons to believe that it is always possible – and, in fact, sometimes necessary – for SIs to be computed in embedded positions. As we mentioned at the very beginning, it is not clear how this possibility could be made consistent with a Gricean approach to SIs. More specifically, the facts suggest that SIs are not pragmatic in nature but arise, instead, as a consequence of semantic or syntactic mechanisms, which we've characterized with the operator, O. This operator, although inspired by Gricean reasoning, must be incorporated to the theory of syntax or semantics, so that – like the overt operator *only* – it will find its way to embedded positions.

In this concluding section we would like to mention a few other observations that support this conclusion. Our discussion will be extremely sketchy given the limited space we have available. Nevertheless, we will try to introduce the pertinent issues and will refer the reader to relevant literature where the basic line of reasoning is developed in greater detail.

4.1. Obligatory Scalar Implicatures

A property that is commonly attributed to SIs is their optionality, sometimes referred to as "cancelability":

(77) a. John did some of the homework. In fact, he did all of it.
b. John did the reading or the homework. He might have done both.

The first sentences in (77)a and (77)b would normally lead to SIs (*not all* and *not both*, respectively). But these SI are not obligatory, else the continuation would lead to a contradiction.

The optionality observed in (77) is a necessary attribute of SIs from the (neo-)Gricean perspective. SIs are not automatic from this perspective, but rather, follow from two assumptions that don't always hold, namely the assumption that the speaker is opinionated about stronger alternatives (which justifies the epistemic step alluded to in section 1.2.), and the assumption that the stronger alternatives are contextually relevant. The fact that these assumptions are not necessarily made in every context explains optionality.

Consider, first, (77)a. This pair of sentences would typically be uttered in a context in which the stronger alternative to the first sentence is not (at least not initially) relevant. For example, the utterance might be a reply to the question: *who did some of the homework?*. And, as a reply to such a question, there is no need to indicate whether anyone did all of the homework, hence no SI is predicted.[38] Consider now (77)b. Here, the stronger alternative is most likely contextually relevant. But, the second sentence indicates that the speaker is not opinionated about this stronger alternative, and the epistemic step is, consequently, not taken.

This optionality is also captured by our grammatical mechanism. Given what we've said up to now, there is nothing that forces the presence of the operator *O* in a sentence containing a scalar item. Optionality is thus predicted, and one can capture the correlation with various contextual considerations, under the standard assumption (discussed in the very beginning of this paper) that such considerations enter into the choice between competing representations (those that contain the operator and those that do not). However, the assumption that the operator *O* is optional is not a *necessary* assumption. One might assume instead that there is an optional process that activates the alternatives of a scalar item, but that once alternatives are active, SIs are obligatory (see Chierchia 2006 for an implementation).

There is thus an interesting difference between the neo-Gricean view of SIs, which necessarily predicts optionality, and the grammatical alternatives, which predicts optionality only under a particular implementation. All things being equal, one might think that there is a slight advantage to the neo-Gricean proposal. Under the neo-Gricean proposal, the fact that SIs are optional is totally predicted, whereas under the grammatical alternative, the prediction depends on further assumptions.

However, this slight advantage will turn into a disadvantage, if it turns out that the putative fact is not real, i.e., if it turns out that for certain scalar items, SIs are obligatorily (rather than optional). Such a claim is, in fact, implicitly present in Krifka (1995) and Spector (2007a), and has been explicity defended in Chierchia (2004, 2006) and Magri (2007). We cannot go over all of the arguments and will narrow our attention to an argument made by Spector (2007a) in the domain of plural morphology.

Consider the contradictory status of the utterance in (78).

(78) #John read (some) books; (in fact) he (might have) read exactly one book.

This contradiction seems to argue that the first sentence *John read (some) books* (on both its variants) is equivalent, based on its basic meaning, to the statement that there is more than one book that John read. However, the putative basic meaning is too strong to account for the semantic consequences of embedding the sentence in downward entailing environments.

Consider the sentences in (79). The interpretation (at least the one that immediately springs to mind) is stronger than what would be predicted under the putative meaning for the first sentence in (78).

(79) a. John didn't read books.
b. I don't think that John read (some) books.

[38] The second sentence introduces information that is not directly relevant to the question being addressed, a fact which – it is plausible to assume – is indicated by *in fact*.

To see this, focus on (79)a. The sentence (under its most natural interpretation) would be false if John read exactly one book. The same point can be made for (79)b, and both points are illustrated by comparison with sentences in which *John read (some) books* is substituted by a sentence that clearly has the putative meaning, namely *John read more than one book*:

(80) a. John didn't read more than one book.
b. I don't think that John read more than one book.

We seem to be facing something close to a paradox. In order to account for the meaning of (78), *John read (some) books* must have a strong meaning, namely that John read *more* than one book, and in order to account for the meaning of (79), it must have a weaker meaning, namely that John read *at least* one book.

Building on suggestions made by Sauerland (2003), Spector argues that the basic meaning is the *at least one* meaning, and that the stronger meaning (i.e. the *more than one* meaning) is an implicature. Explaining how the strong meaning is derived is a rather complicated matter, which we will have to skip in this context. What is important, however, is that implicatures can easily disappear in downward entailing contexts, which accounts for the most natural readings of (79)a and (79)b. As explained in section 2.1., the fact that scalar items need not retain their strengthened meaning in DE contexts is an automatic consequence of the neo-Gricean approach. It is also an automatic consequence of the grammatical perspective that we are advocating, since an exhaustivity operator (even if obligatorily present) need not be inserted below a DE operator. (Under the grammatical perspective, an exhaustivity operator *can* be inserted below a DE operator, a possibility that, as we've seen, is realized in the case of 'intrusive' implicatures.) To see this in detail, let us focus on the case at hand.

If we assume that the plural morpheme *pl* makes it *obligatory* to insert the operator *O* in some syntactic position that c-commands *pl*, we expect the following: in a simple, non-embedded context, *O* can only be inserted just above the plural morpheme, which gives rise to an *at-least two* reading (as demonstrated in Spector's work); but if *pl* occurs in the scope of a DE-operator, there are more choices as to where exactly *O* can be inserted; in particular, *O* may be inserted at the top-most level, i.e. above the DE-operator, in which case no implicature will arise.[39] In the case of the plural morpheme, we therefore predict that the *at least two*-reading *can* disappear in DE-contexts, while it is obligatory in non-embedded UE contexts. This will generally be a property of *obligatory scalar implicatures*: the strengthened meaning of an item that *must* occur under the scope of O will be the only possible one in non-embedded UE contexts, but will appear to be optional in DE contexts. In the case of the plural morpheme, this prediction seems to be correct, since the *at-least-two* reading actually *can* be maintained in DE-contexts, with the appropriate intonation pattern:[40]

[39] Because inserting *O* at the top-most level always gives rise to the reading that is predicted by the pragmatic, neo-Gricean, approach.

[40] The fact that a certain intonation pattern seems to be necessary in these cases, as in other cases of 'intrusive' implicatures, is further evidence for the view that the interpretation of the plural morpheme involves SIs.

(81) Jack may have read *one* book; but I don’t think he has read books.

But if all this is correct, it means that the implicature generated by plural morphology is obligatory (which is why (78) is contradictory in every context). As mentioned, this cannot be captured under neo-Gricean assumptions but can be made to follow from a grammatical theory that incorporates the operator *O*. Specifically, under the grammatical theory, one could claim that for plural morphology, in contrast to standard scalar items, alternatives are automatically activated. Once alternatives are active, they must be associated with the operator *O*, and – to the extent that some alternatives are not entailed by the basic sentence – implicatures are obligatory.[41] This operator yields the *more than one* interpretation for (78).[42] However, once (78) is embedded under a downward entailing operator (e.g. ((79)), the stronger alternatives are now weaker, and the relevant implicatures are eliminated.

This is just one example, and our discussion has been extremely sketchy (as promised). However, we hope that the nature of the argument is clear. Gricean implicatures must be optional/cancelable. But if implicatures are derived by a grammatical mechanism, they are optional only if the mechanism is optional, and that, in turn, is up for grabs. A similar argument has been made in other domains. Most famously, Krifka (1995) has argued that negative polarity items are obligatorily associated with alternatives, and that these alternatives yield obligatory implicatures which account for the distributional properties. This argument has been developed by Chierchia (2004) to account for intervention effects (sketched in section 4.3 below) and has been extended to other polarity items in Chierchia (2006). For the actual details, we will have to refer the reader to the original sources.

4.2. Encapsulation

Consider the oddness of the following:

(82) John has an even number of children. More specifically, he has 3 (children).

The source of the oddness is intuitively clear: the second sentence seems to contradict the first sentence. However, it is not trivial to account for the contradiction. The second sentence *John has 3 children* (henceforth just *3*) has an interpretation which is consistent with the first sentence, e.g., an interpretation consistent with John having exactly 4 children, which is, of course, on even number.[43] So, why should the two sentences feel contradictory? If in the context of (82), *3* was required to convey the information that John has exactly 3 children, the contradiction would be accounted for. But what could force this “exactly” interpretation on the sentence?

[41] If the alternatives are innocently excludable, see fn. 36.

[42] The relevant alternative for Spector is *John read exactly one book*, which is itself generated by appending *O* to yet another sentence.

[43] We are assuming here that the relevant sentences containing numerals have an *at least* interpretation, which may or may not be their only reading. The point we are making here remains valid even if the relevant sentences are ambiguous between an 'at-least' interpretation and an 'exact'-interpretation, as has been suggested by various authors (e.g. Horn 1992, Carston 1998, Geurts, to appear-b), since even under such a view, (82) should have a non-contradictory reading, based on the *at-least* interpretation.

It is tempting to suggest that the theory of SIs should play a role in the account. If in (82) the implicature is obligatory, then the second sentence would contradict the first. And indeed, as we see in (83), and as we've already seen in the previous sub-section, there are some cases where implicatures are obligatory:

(83) Speaker A: Do you know how many children John has?
Speaker B: Yes, he has 4 children. #In fact, he has 5.

However, it turns out that the Gricean reasoning that has been proposed to account for SIs does not derive the attested contradiction. This is a point that was made in a different context by Heim (1991), and was discussed with reference to (82) and similar examples in Fox (2004a) and Magri (2007).[44] To understand the argument, it is useful to try to derive the SI, along the lines outlined in section 1, and to see where things break down.

So, let's try. Upon hearing the utterance of *3*, the addressee (h, for hearer) considers the alternative sentences in (84), and wonders why the speaker, s, did not use them to come up with alternative utterances.

(84) a. More specifically, he has 4 children.
b. More specifically, he has 5 children.
c. …

Since all these (clearly relevant) alternatives are stronger than s's actual utterance, h concludes based on (the assumption that s obeys) the Maxim of Quantity that s does not believe any of these alternatives. i.e., s derives the conclusions in (85), which together with the basic utterance, *3,* can be summarized as (86).

(85) a. $\neg B_s$(John has 4 children).
b. $\neg B_s$(John has 5 children).
c. …

(86) O_{ALT} [B_s(John has 3 children)].

Now, based on the assumption that s is opinionated with respect to the alternatives in (84), h might take 'the epistemic step' (tantamount to 'neg-raising'), which leads to the conclusions in (87), summarized in (88).

(87) a. $B_s\neg$(John has 4 children).

[44] As various examples in Magri (2007) illustrate, our argument does not specifically rely on the use of numerals, which we selected here to simplify the exposition. The general point can be made with other scalar items, even though one has to construct more complicated discourses, as in the following:

(i) Every student, including Jack, solved either none of the problems or all of the problems. #Jack solved some of the problems.

In this case, the second sentence, under its non-strengthened, logical meaning, is compatible with the first (and contextually entails that Jack solved all of the problems). Yet it is felt as contradictory.

So even if numerals only had an 'exact' meaning, as argued by Breheny (to appear), our general point would remain.

b. $B_s\neg$(John has 5 children).
c. …

(88) $B_s[O_{ALT}$ (John has 3 children)].

This conclusion clearly contradicts the first sentence in (82), thus, accounting for the observed phenomenon. We thus seem to have a purely neo-Gricean account of the deviance of (82). But this impression is mistaken, as the following illustrates.

The problem is that we were too quick to derive the conclusions in (85) based on the Maxim of Quantity. It is true that all of the utterances in (84) are *logically* stronger than *3*, but are they all also *more informative*, given the special properties of the immediate context? To answer this question we have to understand what is taken to be true at the point at which *3* is uttered (i.e. after the first sentence in ((82)). If the first sentence in (82) is already taken to be true, i.e. if it is assumed that John has an even number of children, the proposition that John has at least 3 children (the relevant meaning of *3*), and the proposition that John has at least 4 children (the relevant meaning of (84)a) provide *exactly the same information*, namely that John has an even number of children greater or equal to three, i.e., that he has 4 or more children.

So, the Maxim of Quantity does not require s to prefer (84)a to *3*. Therefore, the inference in (85)a does *not* follow from the assumption that s obeys the maxim. Moreover, since (84)a and the second sentence in (82), which uses the number *3*, convey exactly the same information, they are predicted to yield exactly the same SI, which together with the basic contextual meanings amounts to the proposition that John has an even number of children greater or equal to 3, but does not have an even number of children greater or equal to 5, which is, of course, tantamount to saying that John has exactly 4 children. So the only implicature we get by employing this purely Gricean reasoning fails to make (82) incoherent.

In other words, on closer scrutiny, it turns out that we fail to account for the contradictory nature of (82). The Gricean reasoning predicts that (82) will be just as appropriate as the following:

(89) John has an even number of children. More specifically, he has 4 children.

This is in sharp contrast with what happens if SIs are derived within the grammar, using the operator O. Under such a view, the contradiction is derived straightforwardly. The sentence *3* activates alternatives which are operated on by O 'blindly', as it were, and when this happens we obtain the proposition that John has exactly 3 children, and this proposition directly contradicts the earlier sentence which asserts that John has an even number of children.

To couch it differently, what (82) seems to teach us is that the notion of informativity relevant to SI computation is logical entailment, rather than entailment given contextual knowledge. This means that the module that computes SIs has to be encapsulated from contextual knowledge, which makes sense if the module is (part of) grammar but not if it is (part of) a "central system" used for general reasoning about the world, as Grice envisioned. For further arguments to this effect, see Fox and Hackl (2006) and Magri (2007).

4.3. NPIs and Intervention Effects

NPIs are known to be licensed in downward entailing contexts (Ladusaw 1979, Fauconnier 1975a). However, as pointed out by Linebarger 1987, certain logical operators appear to disrupt this licensing:

(90) a. John didn't introduce Mary$_1$ to anyone she$_1$ knows
b. *John didn't introduce [every woman]$_1$ to anyone she$_1$ knows.

This intervention effect has been studied extensively in the literature. Among the important observations that have come out of this study is a typology of the logical operators that yield an intervention effects, henceforth intervening operators (cf., among others, Linebarger 1987, Guerzoni 2000, 2006). Compare (90)b to (91), and (92)a to (92)b.

(91) John didn't introduce [a single woman]$_1$ to anyone she$_1$ knows

(92) a. John didn't talk either to Mary or to any other girl.
b. *I didn't talk both to Mary and to any other girl.
.

This comparison leads to the conclusion that existential quantification and disjunction are not capable of yielding intervention affects, but universal quantification and conjunction are (Guerzoni 2000, 2006). Why should this be the case? Chierchia (2004) suggests that the answer follows from the theory of SIs. We cannot go over the details of the proposal but we can introduce the basic idea. Assume first that licensing of NPIs requires them to be in a DE context. Assume, furthermore, that SIs must be obligatorily added in (90)-(92) (i.e. that we are here in presence of obligatory SIs, just like in the examples considered in section 4.1). It can be shown that while adding the SIs in (90)a, (91) and (92)a retains the DE character of the context, adding the SIs to (90)b and (92)b does not. Thus, in these latter cases, we no longer have DE contexts and hence the condition for the proper licensing of NPIs is not met.

To see how SIs could affect downward entailingness, consider the relationship between the sentences in (93). (93)a seems to entail (93)b, a fact that can be attributed to the presence of negation, a downward entailing operator. But intuitions are less clear for the pair (93)c,d

(93) a. John didn't talk to professors.
b. John didn't talk to physics professors.
c. John didn't talk both to students and to professors
d. John didn't talk both to students and physics professors

Although sentence (93)c does seem to entail (93)d, this is not obvious upon further scrutiny. The reason for this is that (93)c triggers the SI that the stronger alternative with disjunction – (94)a below– is false, namely the implicature that the speaker talked either to students or to professors. If we factor SIs into basic meanings (deriving *strengthened* *meanings*), (93)c can be true while (93)d is false. [To see this, consider a situation where John talked to two biology professors and to no one else. (93)c would be true on its strengthened meaning and (93)b, while true on its weak meaning, would be false on its strengthened meaning.] In other words, there is no entailment between the sentences in (93)c-d, if they receive the syntactic parse in (93)'.

(93)' a. O_{ALT}[John didn't talk both to students and to professors].
b. O_{ALT}[John didn't talk both to students and to physics professors].

The situation in (94) is very different. (94)a has no SIs. The reason for this is simple: the alternative with conjunction – (93)a – is a weaker alternative and is therefore not excluded by any of the approaches to SIs.

(94) a. John didn't talk to students or to professors.
b. John didn't talk to students or to physics professors.

So, even if (94) received a parse with *O*, parallel to (93)', this will not interfere with downward entailingness:

(94)' a. O_{ALT}[John didn't talk both to students or to professors].
b. O_{ALT}[John didn't talk both to students or to physics professors].

O is vacuous in (94)', and therefore does not affect the entailment between the (a) and the (b) sentence. In other words, if we exhaustify, the NPI in (92)b is no longer in a downward entailing environment, while in (92)b downward entailment is not affected.

The same applies, arguably, to all interveners. For example, we can make sense of the fact that universal quantifiers are interveners but existential quantifiers are not. Existential quantifiers, in contrast to universal quantifiers, are the lowest members of their scale. Existential quantifiers and universal quantifiers form a Horn scale in which the universal is the logically strong member. Since, strength is reversed under downward entailing operators, universal quantifiers lead to matrix implicatures when embedded under such operators and existential quantifiers do not. The relevant implicatures, in turn, destroy downward entailingness, thus yielding the intervention effect.

But of course, none of this can work if SI are computed outside grammar. Under such an architecture, there is no reason why they should affect the licensing of NPIs. Moreover, Chierchia shows that his account can work only under very specific assumptions about the effects of SIs on syntactic intervention effects. If there is something to the account, SIs clearly must be computed within grammar.

4.4. Free Choice

An utterance of the sentence in (95) is typically interpreted as a license to choose freely between two available options (the free choice inference, henceforth Free Choice).

(95) You are allowed to eat cake or ice cream.
There is at least one allowed world where you eat cake or ice cream.

More specifically, (95) licenses the two inference in (96).

(96) Free Choice (inference of ((95))
  a. You are allowed to eat cake.
   There is at least one allowed world where you eat cake.
  b. You are allowed to eat ice cream.
   There is at least one allowed world where you eat ice cream.

Free Choice, however, does not follow in any straightforward way from the basic meaning of the sentences. (95) – which contains two logical operators: the existential modal allowed and the disjunction or – should express the proposition that the disjunction holds in at least one of the allowed worlds [$\Diamond(C \vee IC)$]. And, the truth of this proposition does not guarantee that for each disjunct there is an allowed world in which the disjunct is true. [$\Diamond(C \vee IC) \neq> \Diamond(C) \wedge \Diamond(IC)$.]

Kamp (1973), who identified the puzzle, suggested that it be resolved by strengthening the basic semantics of the construction and a solution along such lines has been worked out also in Zimmerman (2000) and Geurts (2005).[45] However, Kratzer and Shimoyama (2002) - henceforth K&S- and Alonso-Ovalle (2005) pointed out that such a revision would get the wrong result when the construction is embedded in a downward entailing environment:

(97) No one is allowed to eat cake or ice cream

If (95) – as part of its basic meaning – were to entail Free Choice, we would expect (97) to be true if one of the free choice inferences in (96) were false for every individual in the domain (e.g. if half the people were allowed to eat cake the other half were allowed to eat ice cream, but no one was free to choose between the two desserts). But (97) seems to express a much stronger proposition, namely that no one is allowed to eat cake and that no one is allowed to eat ice cream.

We've already seen this pattern, namely an inference that appears when a sentence is uttered in isolation, but is not computed as part of the meaning when the sentence is further embedded in a downward entailing environment. We've also seen that this otherwise puzzling pattern would follow straightforwardly if the inference could be derived as an implicature.[46] In the case of Free Choice, K&S suggest that the inference should follow from a reasoning process about the belief state of the speaker that one might call meta-implicature[47].

Specifically, K&S suggest that the sentences in (95) has the alternatives given in (98) below, for which we've argued on independent grounds in section 3.

[45] See also Simons (2005).

[46] To be more precise, as observed in section 4.1., what the grammatical theory of SIs predicts is that the strengthened reading, i.e. in this case the free-choice reading, *can* disappear in DE contexts. In the next section we'll propose an account of why there seems to be a *preference* for the 'literal', non-strengthened reading of scalar items in DE-contexts, i.e. for the markedness of the relevant intrusive implicatures.

[47] Several other works argue that the free-choice inference is an implicature. See in particular Schulz (2005), Klinedinst (2006), Chemla (2008), for various interesting proposals.

(98) Alternatives for (95) proposed by K&S/Alonso-Ovalle:
a. You are allowed to eat the cake.
b. You are allowed to eat the ice cream.

Furthermore, they suggest that when a hearer h interprets s's utterance of (95), h needs to understand why s preferred (95) to the two alternatives. K&S, furthermore, suggest that it is reasonable for h to conclude that s did not choose the alternative because she was not happy with their strong meaning (basic meaning + implicatures) – hence our term meta-implicature. Specifically, K&S suggest that h would attribute s's choice to the belief that the strong meanings of (98)a and (98)b (stated in (99)) are both false.

(99) Strong meaning of the alternatives for (95)
a. You are allowed to eat the cake and you are not allowed to eat the ice cream.
b. You are allowed to eat the ice cream and you are allowed to eat the cake.

And, as the reader can verify, if (95) is true and the strengthened alternatives in (99) are both false, then the Free Choice inferences in (96) would follow.

We believe this logic is basically correct, but we don't see a way to derive it from basic principles of communication (Maxims). In fact, if s believed that the Free Choice inferences hold, the Maxim of Quantity would have forced s to prefer the sentences in (98), and to avoid an utterance of (95) altogether. The fact that s did not avoid (95) should, therefore, lead h to the conclusion that s does not believe that the sentences in (98) are true (see our discussion in section 1.1.).

This has led Chierchia (2006) and Fox (2007) to provide a formal/grammatical alternative to K&S. We cannot go over the details of the proposals, but would like to point out Fox's observation, namely that K&S's results follow from a representation in which two instances of the operator, *O*, are appended to (95):

(100) A logical form for (95) that derives Free Choice:
OO(You are allowed to eat cake or ice cream).
There is at least one allowed world where you eat cake or ice cream. And (99)a,b, are both false.

Furthermore, we would like to refer the reader to Chierchia (2007, SALT) where constraints on the relevant grammatical representations yields an account of the cross-linguistic distribution of Free Choice items.

In conclusion, we have sketched reasons to believe that free choice effects can be explained in a principled way as meta- (or higher order) implicatures. If this is anywhere close to the mark, then clearly implicatures must be part of grammar.

4.5. Non-monotonic contexts: negating alternatives that are neither stronger nor weaker.

Consider the following sentence:

(101) Exactly one student solved some of the problems

Let's assume that (101)'s only scalar alternative is (102).

(102) Exactly one student solved all of the problems

(102) is neither stronger nor weaker than (101): both of them can be true in a situation where the other is false. Since (102) is not more informative than (101), Grice's maxim of quantity, under its most natural understanding, does not require that one utter (102) rather than (101) in case both are believed to be true and relevant. So the Gricean approach, unless supplemented with quite specific assumptions, predicts no SI in the case of (101). In contrast to this, a theory that incorporates the exhaustivity operator[48], which is modeled on the semantics of *only*, does predict an implicature for (101). Indeed, applying the exhaustivity operator to a given sentence S with alternatives ALT(S) generally returns the conjunction of S and of the negations of all the alternatives of S that are not entailed by S[49], which include both alternatives that are stronger than S and possibly alternatives that are neither stronger nor weaker than S. So the strengthened meaning of (101) is predicted to be the proposition expressed in (103)a, which is equivalent to (103)b:

(103) a. Exactly one student solved some of the problems and it is false that exactly one student solved all of the problems
b. One student x solved some of the problems, x did not solve all of the problems, and none of the other students solved any of the problems.

It seems to us that this prediction is borne out: (103) is indeed a very natural interpretation for (101). The mere fact that implicature computation seems to involve the negation of non-stronger alternatives is quite unexpected from the Gricean perspective.[50]

4.6. Constraints on the placement of the exhaustivity operator: a preference for stronger interpretations

We have observed that a hallmark of SIs is that they tend to disappear in downward-entailing environments – i.e. the strengthened reading of scalar items is dispreferred under, say, negation or in the restrictor of a universal quantifier. At first sight, this phenomenon makes the pragmatic, neo-gricean, account of SIs particularly appealing:

[48] See fn. 8 and fn. 36.

[49] This is a simplification, given the modifications that we adopted above in order to reach a correct treatment of disjunctive sentences. See fn. 36 .

[50] Van Rooij & Schulz (2004, 2006) and Spector (2003, 2006, 2007b) show that this fact could however be made to follow from a purely pragmatic approach if it is assumed that the alternatives of a given sentence are always closed under conjunction. These works aim at deriving the exhaustive interpretation of answers to wh-questions in a Gricean way. They assume that the alternatives of a given *positive* answer to a wh-question consist of the set of all positive answers, which is closed under conjunction. While this might make sense in the context of question-answer pairs, it is by no means obvious that it can naturally be extended to all cases of SIs.Furtermore, Fox's (2007) account of free-choice effects, need to assume that the alternatives of a sentence are not always closed under conjunction.

indeed, as we have seen, the absence of the strengthened reading in DE contexts is directly predicted from the neo-gricean perspective. However, as we pointed out in section 2, the strengthened meaning of a scalar item is actually not ruled out in DE contexts; it is only dispreferred. From a purely Gricean perspective, it is a challenge to explain why a scalar item could ever be interpreted under its strengthened meaning in a DE context (so called 'intrusive implicatures'). To account for such cases, advocates of the purely pragmatic perspective are forced to introduce new mechanisms (but if our previous arguments are conclusive, these 'repairs' are anyway unable to account for the full range of phenomena).[51] The grammatical view does not face a similar challenge; but it clearly needs to be supplemented with some principles that determine which particular readings are preferred and which ones are dispreferred (and hence marked).

One possibility that suggests itself is that, when a sentence is potentially ambiguous, there is a preference for the strongest possible interpretation. Such a general principle has been suggested independently by various researchers beginning with Dalrymple, Kanazawa et al. (1998) – the 'strongest meaning hypothesis'. If a principle of this sort is adopted, then inserting *O* in a DE context would be dispreferred: indeed, for any sentence S, O(S) is stronger than S; hence, inserting O(S) in the (immediate) scope of a DE operator X, i.e. an operator that reverses logical strength, gives rise to a sentence X(O(S)) that is now *weaker* than what would have resulted if O were absent, i.e. weaker than X(S).

How exactly such a principle should be stated is far from trivial. We will briefly mention two possible implementations, and point out two cases where they make different predictions. A conceivable version of the principle is the following:

(104) **Strongest Meaning Hypothesis (Global Version)**
Let φ be a certain logical form. Let φ's competitors be all the LFs that differ from φ only with respect to where exhaustivity operators occur. Then, everything else being equal, φ is dispreferred if one of its competitors is stronger than φ.

Such a principle predicts that the preferred reading is always the *strongest possible one* (if there is one) among all the possible readings. Note that such a principle does not rule out any reading, but only predicts that some readings will be preferred. Yet there might be many contextual reasons why the interpreter of a sentence could choose a reading that is not the strongest one but is still relatively highly ranked (in the sense that it is stronger than many other possible readings). Nevertheless, we expect that the weakest readings among all the possible readings will be felt as 'marked' – this will be in particular the case with the readings that involve an embedded implicature in a DE-environment.

Such a principle might be criticized on the ground that it involves quite a heavy computational load – the number of LFs to be compared in order to apply (104) to a given LF will be roughly equal to 2^n, where n is the number of sites where *O* can be inserted. So we might consider another version of the 'strongest interpretation principle',

[51] Horn (1989), for instance, resorts to the notion of metalinguistic negation, as mentioned in section 2.2., which might be generalized to other operators, while others (Levinson 2000, Geurts to appear-a) must acknowledge the existence of embedded implicatures, which they however view as a peripheral phenomenon.

one which would be more ‘local’. Specifically, one might suggest hat each occurrence of the exhaustivity operator should be unmarked if it gives rise to a reading that is stronger than what would have resulted in its absence. Such a principle can be stated as follows:

(105) **Strongest Meaning Hypothesis (Local Version)**
Let S be a sentence of the form $[_{S}\ldots\ldots O(X)\ldots\ldots]$. Let S' be the sentence of the form $[_{S'}\ \ldots\ldots X\ \ldots\ldots]$, i.e. the one that is derived from S by replacing O(X) by X, i.e. by eliminating this particular occurrence of O. Then, everything else being equal, S' is preferred to S if S’ is logically stronger than S.

Such a principle, like the previous one, predicts that O should be dispreferred under DE-operators. Yet (104) and (105) do not make exactly the same predictions, as we now illustrate with two particular cases. First consider the following:

(106) For this class, we must read most of the books on the reading list

An exhaustivity operator could be inserted either above or below the modal *must*, giving rise to the following readings :

(107) a. O(we must read most of the books on the reading list)
= we must read most of the books on the reading list and we don’t have to read all of then
b. We must O(read most of the books on the reading list)
= we must read most of the books on the reading list and we have the obligation no to read them all

Clearly, (107)b strictly entails (107)a, and is therefore predicted to be preferred by the principle stated in(104). In contrast, (105) makes no such prediction. This is so because according to (105), (107)a and (107)b are not competitors of each other. Rather, each of them is to be compared to the proposition that one gets be deleting the operator, namely to the non-strengthened reading of ‘We must read *Ulysses* or *Madame Bovary*”. Plainly, the condition stated in (105) is met in both cases, since in both cases the presence of O has a strengthening effect. In fact, in UE contexts, the principle in (105) does not generally favor one particular insertion site for the exhaustivity operator. Of course, more general considerations (such as, for instance, the plausibility of a given reading) might create a preference for certain readings.

At first sight, it might seem that the principle in (105) is to be preferred to the one in (104), because (107)b seems to be a much less natural reading than (107)a (while (104) predicts the reverse). But such a conclusion is not warranted, because the preference for the strongest interpretation stated in (104) can anyway be overridden by other considerations, such as, for instance, plausibility considerations (it is very unlikely from the start that we are forbidden to read all the books on the reading list).

Another case where the two principles in (104) and (105) make different predictions is provided by non-monotonic contexts. Consider again the following sentence:

(108) Exactly one student solved some of the problems

Here are the three possible parses for (108):

(109) a. Exactly one student solved some of the problems <no exhaustivity operator>
b. O(Exactly one student solved some of the problems)
c. Exactly one student O(solved some of the problems)

As explained in the previous section, the predicted reading for (109)b is (110)a, which is in turn equivalent to (110)b:

(110) a. Exactly one student solved some of the problems, and it is false that exactly one student solved some of the problems
b. There is only one student who solved any of the problems, and that student didn't solve all of the problems

On the other hand, (109)c is equivalent to the following:

(111) Exactly one student solved some, but not all of the problems.

Now, it turns out that the proposition expressed in (110)a,b is strictly stronger than (111). Indeed, any situation where (110)a,b is true is one in which (111) is true. But the reverse entailment does not hold: for in a situation where one student solved some, but not all, of the problems, while all the other students solved all of the problems, (111) is true but (110) is false. Therefore the principle stated in (104), according to which the (globally) strongest interpretation should be preferred, predicts that the parse in (109)c should be dispreferred (since it expresses a reading that is weaker than the one expressed by (109)b). More generally, a principle such as (104) tends to disfavor embedded implicatures in non-monotonic contexts. On the other hand, on the more 'local' principle stated in (105), both (109)b and (109)c are predicted to be fine, since none of them is logically weaker than what results from deleting the exhaustivity operator, i.e. (109)a ((109)b strictly entails (109)a, while (109)c is neither weaker nor stronger than (109)a).

We are not going to assess here which predictions are closer to the facts; nor are we going to investigate other conceivable implementations of the strongest interpretation hypothesis. Our goal in this section was just to hint at the kind of issues that arise once we adopt the grammatical perspective on SIs that we have advocated in this paper.

4.7. Concluding remarks

In this paper we tried to show that SIs can occur in all sorts of embedded context. If this attempt has been successful, we think it calls for a reassessment of the semantics/pragmatics interface. In order to establish our point, we have adopted the view that implicatures arise through a silent exhaustification operator, akin to *only*, which acts on scalar alternatives. While this has not been a crucial ingredient of what we've done (one can imagine alternative ways of stating the point about embedding), we think that the idea – while leaving many open issues – has significant benefits: in many cases (involving Hurford's Constraint, iterated applications of *O*, etc.) it makes just the right predictions and no viable alternative seems to be in sight.

The grammatical view of SIs retains the most beautiful feature of the Gricean insight: the sensitivity of SIs to embeddings within polarity affecting contexts. And, through the link to alternative sensitive operators, creates a powerful bridge to a host of like phenomena occurring in very diverse corners of grammars (from the analysis of plurals, throuh free choice, to intervention and the like). Within the limits of the present paper, these remain largely promissory notes. But we hope that we were able to lay out the strategy that needs to be pursued in a fairly clear manner. Finally, we hope that it will be possible begin to reap the benefits of the entrance of SIs (and of, possibly, implicatures of other sorts) into the computational system of grammar.

References

Alonso-Ovalle, L. (2005). Distributing the Disjuncts over the Modal Space. NELS 35, University of Massachusetts, Amherst, GLSA.

Atlas, J. D. and S. Levinson (1981). It-Clefts, Informativeness, and Logical Form: Radical Pragmatics (Revised Standard Version). Radical Pragmatics. P. Cole. New York, Academic Press**:** 1-61.

Bach, K. (1994). "Conversational Impliciture." Mind & Language **9**(2): 124-162.

Bach, K. and R. M. Harnish (1979). Linguistic Communication and Speech Acts. Cambridge, Mass., MIT Press.

Breheny, R. (to appear). "A New Look at the Semantics and Pragmatics of Numerically Quantified Noun Phrases." Journal of Semantics.

Carnap, R. (1950). Logical Foundations of Probability. Chicago, University of Chicago Press.

Carston, R. (1988). Implicature, Explicature, and Truth-Theoretic Semantics. Mental Representations: The Interface Between Language and Reality. R. Kempson. Cambridge, Cambridge University Press**:** 155-181.

Carston, R. (1998). Informativeness, relevance and scalar implicature. Relevance theory: applications and implications. R. Carston and S. Uchida. Amsterdam, Benjamins**:** 179-236.

Chemla, E. (2008). "Similarity: Towards a Unified Account of Scalar Implicatures, Free Choice Permission and Presupposition Projection." Unpublished Article. http://www.emmanuel.chemla.free.fr/Material/Chemla-SIandPres.pdf.

Chierchia, G. (2004). Scalar Implicatures, Polarity Phenomena, and the Syntax/Pragmatics Interface. Structures and beyond. A. Belletti. Oxford, Oxford University Press. **3**.

Chierchia, G. (2006). "Broaden your Views. Implicatures of Domain Widening and the "Logicality" of Language." Linguistic Inquiry **37**(4): 535-590.

Dalrymple, M., M. Kanazawa, et al. (1998). "Reciprocal expressions and the concept of reciprocity." Linguistics and Philosophy **21**: 159-210.

Davis, W. (1998). Implicature: Intention, Convention, and Principle in the Failure of Gricean Theory. Cambridge, Cambridge University Press.

Ducrot, O. (1973). La preuve et le dire. Paris, Mame.

Fauconnier, G. (1975a). "Polarity and the Scale Principle." Chicago Linguistics Society **11**: 188-199.

Fauconnier, G. (1975b). "Pragmatic Scales and Logical Structure." Linguistic Inquiry **VI**(3): 353-376.

Fox, D. (2004a). "Implicatures and Exhaustivity - Class 4: Back to the Theory of Implicatures." Lecture notes for a class taught in November 2004 at U.S.C. http://web.mit.edu/linguistics/people/faculty/fox/class_4.pdf.

Fox, D. (2004b). "Implicatures and Exhaustivity - Class 5: *Only* a little bit more." Lecture notes for a class taught in November 2004 at U.S.C. http://web.mit.edu/linguistics/people/faculty/fox/class_5.pdf.

Fox, D. (2007). Free Choice Disjunction and the Theory of Scalar Implicatures. Presupposition and Implicature in Compositional Semantics. U. Sauerland and P. Stateva. New York, Palgrave Macmillan**:** 71-120.

Fox, D. and M. Hackl (2006). "The Universal Density of Measurement." Linguistics and Philosophy **29** (5): 537-586.

Gamut, L. T. F. (1991). Logic, Langauge, and Meaning. Chicago, Chicago University Press.

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

Geurts, B. (2005). "Entertaining Alternatives: Disjunctions as Modals." Natural Language Semantics **13**(4): 383–410.

Geurts, B. (to appear-a). "Scalar Implicature and Local Pragmatics." Mind and Language.

Geurts, B. (to appear-b). Take "five": the meaning and use of a number word. Indefiniteness and plurality. T. Liliane and S. Vogeleer. Amsterdam, Benjamins.

Grice, P. (1989). Studies in the Way of Words. Cambridge, Mass., Harvard University Press.

Groenendijk, J. and M. Stokhof (1984). Studies on the Semantics of Questions and the Pragmatics of Answers, University of Amsterdam.

Groenendijk, J. and M. Stokhof (1990). "Partitioning Logical Space."

Guerzoni, E. (2000). Towards a movement-based account of the locality constraints on Negative Polarity Items. CONSOLVE VIII, SOLE.

Guerzoni, E. (2004). "Even-{NPI}s in Yes/No Questions." Natural Language Semantics **12**(4): 319–343.

Guerzoni, E. (2006). "Intervention Effects on NPIs and Feature Movement: Towards a Unified Account of Intervention." Natural Language Semantics **14**(4): 359-398.

Heim, I. (1991). Artikel und Definitheit. Semantik: Ein internationales Handbuch der zeitgenössischen Forschung. A. v. Stechow and D. Wunderlich. Berlin, Walter de Gruyter**:** 487-535.

Hirschberg, J. (1985). A Theory of Scalar Implicature, University of Pennsylvania.

Horn, L. (1972). On the Semantic Properties of Logical Operators in English, UCLA.

Horn, L. (1985). "Metalinguistic Negation and Pragmatic Ambiguity." Language **61**: 121-174.

Horn, L. (1989). A Natural History of Negation. Chicago, University of Chicago Press.

Horn, L. (1992). "The Said and the Unsaid." SALT **2**: 163-192.

Horn, L. (2005). The Border Wars: a neo-Gricean perspective. Where Semantics meets Pragmatics. T. Turner and K. von Heusinger, Elsevier.

Hurford, J. R. (1974). "Exclusive or Inclusive Disjunction " Foundation of Language **11**: 409-411.

Katzir, R. (2007). Structurally-Defined Alternatives. MIT.

Klinedinst, N. (2006). Plurality and Possibility. PhD dissertation, UCLA.

Kratzer, A. and J. Shimoyama (2002). Indeterminate pronouns: The view from Japanese, Tokyo, Hituzi Syobo.

Krifka, M. (1993). "Focus and Presupposition in Dynamic Interpretation." Journal of Semantics **10**: 269-300.

Krifka, M. (1995). "The Semantics and Pragmatics of Polarity Items." Linguistic Analysis **25**: 209-257.

Kroch, A. (1972). "Lexical and Inferred Meanings for Some Time Adverbs." Quarterly Progress Reports of the Research Laboratory of Electronics **104**: 260-267.

Ladusaw, B. (1979). Polarity Sensitivity as Inherent Scope Relations, University of Texas at Austin.

Landman, F. (1998). "Plurals and Maximalization." **Events and Grammar**: 237–272.

Larson, R. (1985). "On the Syntax of Disjunction Scope." Natural Language and Linguistic Theory **3**: 217-264.

Levinson, S. (2000). Presumptive Meaning. Cambridge, Mass., MIT Press.

Levinson, S. C. (1983). Pragmatics. Cambridge, U.K., Cambridge University Press.

Lewis, D. (1973). Counterfactuals. Oxford, Blackwell.

Linebarger, M. (1987). "Negative Polarity and Grammatical Representation." Linguistics and Philosophy **10**: 325-387.

Magri, G. (2007). "A Theory of Individual Level Predicates Based on Blind Scalar Implicatures." Unpublished Paper. M.I.T.

Matsumoto, Y. (1995). "The Conversational Condition on Horn Scales." Linguistics and Philosophy **18**(1): 21-60.

Récanati, F. (2003). "Embedded implicatures." Philosophical Perspectives(17): 1299-1332.

Romero, M. and C.-H. Han (2004). "On Negative Yes/No Questions." Linguistics and Philosophy **27**(5): 609–658.

Rooth, M. (1985). Association with Focus, University of Massachusetts at Amherst.

Rooth, M. (1992). "A theory of focus interpretation." Natural Language Semantics **1**(1): 117-121.

Russell, B. (2006). "Against Grammatical Computation of Scalar Implicatures
" Journal of Semantics **23**: 361 - 382.

Sauerland, U. (2003). A New Semantics for Number, Cornell University, Ithaca, NY, CLC Publications.

Sauerland, U. (2004). "Scalar Implicatures in Complex Sentences." Linguistics and Philosophy **27**(3): 367–391.

Sauerland, U. (2005). The Epistemic Step. Experimental Pragmatics
Cambridge University, Cambridge, UK. .

Schulz, K. (2005). "A pragmatic solution for the paradox of free choice permission." Synthese: 343-377.

Sharvit, Y. and J. Gajewski (2007). On the Calculation of Local Implicatures. WCCFL 2007.

Simons, M. (2005). "Dividing Things Up: the Semantics of *or* and the Modal/*or* Interaction." Natural Language Semantics **13**: 271-316.

Singh, R. (2006). Eager for Distinctness. Eleventh ESSLLI Student Session

J. Huitink and S. Katrenko**:** 76-89.

Soames, S. (1982). "How Presuppositions are Inherited: A Solution to the Projection Problem." Linguistic Inquiry **13**: 483-545.

Spector, B. (2003). Scalar Implicatures: Exhaustivity and Gricean Reasoning. ESSLLI 2003 Student Session, Vienna.

Spector, B. (2006). Aspects de la pragmatique des opérateurs logiques. PhD dissertation, Université Paris 7.

Spector, B. (2007a). Aspects of the Pragmatics of Plural Morphology. Presupposition and Implicature in Compositional Semantics. U. Sauerland and P. Stateva, Palgrave-Macmillan**:** 243-281.

Spector, B. (2007b). Scalar Implicatures: Exhaustivity and Gricean Reasoning. Questions in Dynamic Semantics. M. Aloni, A. Butler and P. Dekker, Elsevier**:** 225-249.

Sperber, D. and D. Wilson (1986). Relevance: Communication and Cognition. Oxford, Blackwell.

Stalnaker, R. (1968). A Theory of Conditionals. Studies in Logical Theory. N. Rescher. Oxford, Blackwell**:** 98-112.

van Rooij, R. and K. Schulz (2004). "Exhaustive Interpretation of Complex Sentences." Journal of Logic, Language and Information **13**(4): 491-519.

van Rooij, R. and K. Schulz (2006). "Pragmatic Meaning and Non-Monotonic Reasoning: The Case of Exhaustive Interpretation." Linguistics and Philosophy **29**(2): 205–250.

Zimmerman, T. E. (2000). "Free choice disjunction and epistemic possibility." Natural Language Semantics **8**(4): 255-290.