Experimental Evidence for Embedded Scalar Implicatures*

EMMANUEL CHEMLA
Institut Jean-Nicod & LSCP (CNRS, EHESS; DEC, ENS, Paris, France)
BENJAMIN SPECTOR
Institut Jean-Nicod (CNRS, EHESS; DEC, ENS, Paris, France)

December 13, 2010

Abstract

Scalar implicatures are traditionally viewed as *pragmatic* inferences which result from a reasoning about speakers' communicative intentions (Grice 1989). This view has been challenged in recent years by theories which propose that scalar implicatures are a *grammatical* phenomenon. Such theories claim that scalar implicatures can be computed in embedded positions and enter into the recursive computation of meaning—something that is not expected under the traditional, pragmatic view. Recently, Geurts and Pouscoulous (2009) presented an experimental study in which embedded scalar implicatures were not detected. Using a novel version of the truth value judgment tasks, we provide evidence that subjects sometimes compute embedded scalar implicatures.

Keywords: scalar implicatures, localism, globalism, experiment, pragmatics.

1 Theories of scalar implicatures

Scalar implicatures (SIs for short) are usually viewed as *conversational* implicatures (Grice 1967) , i.e. inferences that are not directly encoded in the conventional meaning of the relevant sentences, but rather result from a pragmatic reasoning about the speaker's communicative intentions. According to the neo-Gricean approach to SIs, given a sentence S and a set of competitors for S, called its *scalar alternatives*, the SIs triggered by S should follow from the assumption that the author of S, by choosing S rather than any of its scalar alternatives, complied with Grice's conversational maxims. For instance, the fact that a sentence such as (1) below tends to trigger the inference that John did not solve all of the problems is accounted for in the following manner: it follows from Grice's maxims of conversation that the author of (1) does not consider its more informative alternative ('John solved all of the problems') to be true, for otherwise she should have used this alternative instead (according to the so-called maxims of Quality and Quantity); hence she must consider it to be false. Importantly, this strengthening of 'some' into 'some but not all' is viewed as resulting from a reasoning about a full *speech act*.

(1) John solved some of the problems.

*A shorter presentation of this work containing in particular less precise discussions of the interpretation of the experimental task and fewer results can be found in Chemla and Spector (2010).

The Gricean, pragmatic approach to various types of inferences was challenged nearly as soon as it was presented (see e.g., Cohen 1971), but it nevertheless became the dominant view. In the case of scalar implicatures, however, several works in the last decade (see Landman 1998, Chierchia 2004, Fox 2007, Chierchia, Fox, and Spector in press) have proposed an alternative view, according to which the computation of SIs is a *grammatical* phenomenon, i.e. does not rely on a general reasoning about speakers' intentions, but belongs to compositional semantics. According to this alternative view, SIs are not properties of speech acts, but of linguistic expressions. A consequence of this 'grammatical' approach to scalar implicatures is that, in principle, the mechanism whereby the meaning of a simple sentence S is enriched with scalar implicatures should be able to apply to S even when S is embedded in a more complex sentence. For instance, if the grammatical approach is correct, then the strengthening of 'some' into 'some but not all' could occur under the scope of linguistic operators. In other words, such grammatical approaches allow for *local* enrichment, which is why they are often referred to as *localist* approaches—as opposed to more traditional, so-called *globalist* approaches, in which the computation of scalar implicatures is a process that applies to full speech acts.

Consider the following example:

(2) Every student solved some of the problems.

If we consider the alternative where 'some' is replaced with 'all', we obtain (3):

(3) Every student solved all of the problems.

Since (3) asymmetrically entails (2), standard Gricean reasoning leads to the conclusion that the author of (2) does not believe (3) to be true, for otherwise she should have said so (Grice's maxim of Quantity). With the auxiliary assumption that the speaker is 'opinionated'—i.e. has an opinion as to the truth-value of (3), cf. Spector (2003), Sauerland (2004, 2005), van Rooij and Schulz (2004)—it is predicted that (2) should be interpreted as implying the negation of (3), hence as equivalent to (4):

(4) Every student solved some of the problems and not every student solved all of the problems.

If, however, the strengthening of 'some' into 'some but not all' can occur at an embedded level, one expects that a possible reading for (2) is the one expressed by (5) below:

(5) Every student solved some but not all the problems.

Several recent theories claim that the computation of SIs can occur at an embedded level (Landman 1998, Chierchia 2004, Recanati 2003, Fox 2007 and Chierchia et al. in press). All these localist theories predict that (5) is indeed a possible reading for (2).

Thus it would seem that determining the possible readings of sentences like (2) would provide decisive evidence in the debate between localism and globalism. If, on the one hand, the reading expressed in (5) is indeed a possible reading for (2), then localist theories would be vindicated; if, on the other hand, (5) is not a possible reading for (2), then localist theories would be refuted.

Things are, unfortunately, more complicated, because most current formalized *globalist* theories of SIs (e.g., Spector 2003, van Rooij and Schulz 2004, Sauerland 2004, Chemla 2008, 2009b) also predict (5) to be a possible reading of (2). These theories can derive this reading not by localist means, of course, but by adding to the list of negated scalar alternatives of (2) the proposition: 'Some students solved all the problems'.[1]

As a result, we can now identify three types of theories:

T1. The restricted globalist approach, which predicts that (2) can be interpreted as (4) and cannot be interpreted as (5).

T2. The localist approach, which predicts that (2) can be interpreted as (5).

T3. The non-restricted globalist approach, which predicts that (2) can be interpreted as (5).

Recently, Geurts and Pouscoulous (2008, 2009) presented experimental evidence that they interpret as showing that (5) (for which we will henceforth use the descriptive label 'local' reading) is not a possible reading for (2). If they were right, they would have provided important arguments against theories of type T2 and T3. Such data are therefore crucial in order to assess the on-going debate about the status of scalar implicatures. The first goal of this paper is to provide new experimental data which show, contra Geurts and Pouscoulous' interpretation of their results, that (5) is a possible reading for (2).

However, as we pointed out above, the existence of this reading does not as such settle the debate between localist theories and globalist theories, given the existence of theories of type T3. Hence, a second goal of this paper is to collect experimental data for a case where theories of type T2 and T3 make different predictions.

We will start with a discussion of Geurts and Pouscoulous' study, and point out what we believe are potential limitations of their methodology (section 2). Then we will present our own experimental design, which is intended to overcome these limitations (section 3).

[1] In such theories, the alternatives that are negated do not have to be stronger than the sentence uttered. More specifically, these theories assume that 1) alternatives can be obtained by replacing *several* scalar items of the sentence at the same time (to obtain the alternative above, 'every' is replaced with 'some' and 'some' is replaced with 'all') and 2) the scalar implicatures of a sentence S are derived by negating as many alternatives as possible (i.e. without deriving a contradiction)—that is, by negating not only alternatives that are stronger than S, but also alternatives that are logically independent of S. These assumptions can be motivated from a neo-Gricean perspective, when properly formalized. See Spector (2006), and Magri (2009) for related issues, and Fox (2007) for potential concerns with systematically allowing simultaneuous replacements of scalar items.

In section 4, we will report a first experiment whose results go directly against Geurts and Pouscoulous' conclusions regarding examples such as (2). Finally, in section 5, we will present the results of a second experiment, in which we tested cases that are in principle able to distinguish between theories of type T2 and T3.

2 Geurts and Pouscoulous' results

Geurts and Pouscoulous showed that embedded implicatures are not detected by naive speakers in a variety of experimental settings (Geurts and Pouscoulous 2008, 2009). In this section, we discuss one of their results in particular, namely, the fact that, in a sentence-picture matching task, subjects do not seem to detect the local reading for sentences like (2) (i.e. the reading expressed by (5)). Subsequently, by modifying certain aspects of their original experiment, we will show that the local reading can be detected.

2.1 Description of their results

Geurts and Pouscoulous presented sentences containing the scalar item 'some' in the scope of a universal quantifier, such as (6):

(6) All the squares are connected with some of the circles.

They collected truth-value judgments from naive speakers for such sentences in various situations.

Before describing their results, let us fix some terminology. When we mention the *literal reading* of (6), we mean the reading that results from standard compositional semantics, hence a reading that does not include any SI (this reading is paraphrased in (7a)). The *global reading* of (6) is the reading predicted by theories of type T1 (paraphrased in (7b)). Finally, the *local reading* is the reading predicted by localist theories (theories of type T2) and theories of type T3, in which (6) is interpreted as if 'some' meant 'some but not all' (cf. (7c)).

(7) a. Literal Reading. Every square is connected with at least one circle.[2]
b. Global Reading. Every square is connected with at least one circle, and it is not the case that every square is connected with all the circles.
c. Local Reading. Every square is connected with at least one circle, and no square is connected with all the circles.

[2]We ignore here the contribution of the plural morpheme. This is immaterial given that the experimental items used by Geurts & Pouscoulous contained no square that was connected with exactly one circle. This point also holds for our own experimental items.

Note that the local reading asymmetrically entails the global reading, which itself asymmetrically entails the literal reading.

The crucial condition in Geurts and Pouscoulous' (2009) experiment is the one where subjects are presented with a picture which makes the global reading true but the local reading false. If subjects interpret the target sentence as equivalent to the local reading, they should judge the sentence to be false; if not, they should judge it to be true. Geurts and Pouscoulous used pictures like the one shown in Fig. 1. They reported that all par-

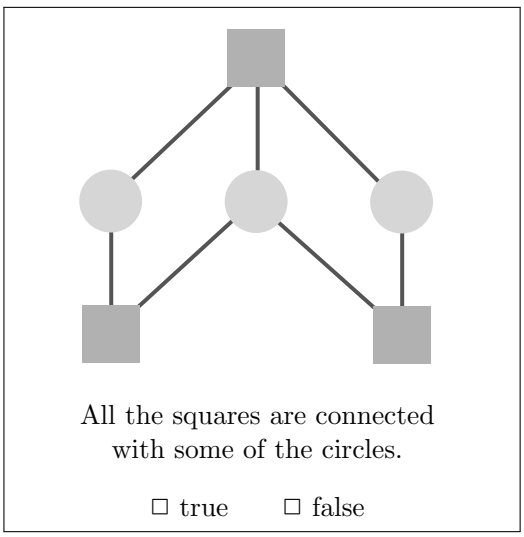


Figure 1: Item from Geurts and Pouscoulous' 2009 experiment 3 (Fig. 2 therein)

ticipants judged the sentence true in this situation, even though the local reading is false (see section 2.2.1 for details). Geurts and Pouscoulous were cautious not to jump to the conclusion that the local reading does not exist. They noted that this result would also follow if the sentence were ambiguous between the global and the local readings, and if the global reading were for some reason preferred in such a forced-choice setting. More specifically, they distinguished between a 'strong' version of localism and a 'weak' version. The strong version, which they call 'mainstream conventionalism' (and is attributed to Chierchia 2004), holds that embedded scalar implicatures are computed *by default*; the weak version, which they call 'minimal conventionalism' (and which can be attributed, e.g, to Chierchia 2006, Fox 2007, Chierchia et al. in press, Magri 2009), only claims that embedded SIs are possible, but not necessarily preferred, and that complex sentences are thus multiply ambiguous, depending on whether an embedded scalar implicature is computed at a given syntactic site.[3] In order to address 'minimal conventionalism', they ran

[3]Geurts and Pouscoulous' 2009 labels seem to us to be potentially misleading, for it is not clear that what they call 'mainstream conventionalism' is currently the dominant view among advocates of the localist

a follow-up experiment in which participants were given the additional option to report explicitly that a sentence is ambiguous by answering 'The sentence could be either true or false'. Even though participants were able to recognize other kinds of ambiguities, none of them reported that the target sentence (6) could be either true or false in a situation such as the one depicted in Fig. 1. Geurts and Pouscoulous interpreted these results as showing that the local reading is not a possible construal of the sentence, from which they concluded that even 'minimal conventionalism' is wrong.

2.2 Potential methodological problems

The outcome of Geurts and Pouscoulous' 2009 experiment is that they did not detect a particular reading. We agree that this result casts doubt on the view that the local reading is the default reading, for if this were so, one would expect that at least some subjects would detect it in such a truth-value judgment task. However, the failure to detect a particular reading in a particular experimental setting (or various experimental settings) does not prove that the reading in question is not one of the possible readings of the relevant sentence. As we will see, there are a number of reasons which may explain why their methodology failed to detect the local reading, even if it existed. We thus object to the stronger claim that Geurts and Pouscoulous (2009) make, i.e. the claim that embedded scalar implicatures do not exist.

In this section, we present what we view as possible limitations of Geurts and Pouscoulous' methodology. The experiments that we will present in the subsequent sections were designed to overcome these limitations.

2.2.1 Salience and Readability

First, we find Geurts and Pouscoulous' (2009) pictures rather difficult to decipher. Consider the example depicted in Fig. 1 again. The crucial bit of information for the present purposes is that the square on top of the picture is connected with all the circles, hereby falsifying the local reading. On purely introspective grounds, we find this information pretty hard to extract, and participants may either miss it or ignore it altogether. In order to assess the truth-value of the local reading, one needs to check for each square, whether it is connected with some, all or none of the circles. This is not a very engaging task, and subjects are likely to avoid it if they can. If the relevant sentence is ambiguous between the literal, the global and the local readings, subjects could thus choose to ignore the local reading.

approach. In particular, Chierchia et al. (in press) make no specific prediction as to whether the local reading is the preferred reading for (2), contrary to what Geurts and Pouscoulous suggest (cf. section 4.6. of Chierchia et al. in press).

2.2.2 Relevance and Disambiguation

We speculate that another reason why the local reading might have been particularly hard to detect in Geurts and Pouscoulous' task, even if it existed, is that the pictures they used failed to make the local reading sufficiently *relevant*. Under a reasonable notion of relevance (such as the one based on Groenendijk and Stokhof's 1984 partition semantics for questions), the local reading ('every square is connected with some of the circles and not with all of them') is relevant typically in a context in which we are interested in knowing, for each square, whether it is connected with some, all, or no circle. Such a context would for instance result from raising the following question: 'Which squares are connected to which circles?'.[4] Even though we do not know how considerations of relevance affect the subjects' performances in a sentence-picture matching task, we might expect that the local reading would be significantly more accessible if we used pictures that prompted subjects to pay attention to the specific properties of each particular square, rather than to more global patterns. We may thus hope that by constructing other sentence-picture matching tasks in which the relevant items are more clearly individuated, we will be able to make the local reading more relevant.

2.2.3 'Preference for Truth'

Various authors have discussed principles which would lead participants to show a preference for the logically weakest reading of a sentence, or even to fail to be aware of the existence of a particular reading R1 when R1 entails another clearly available reading R2. For instance, a principle of 'preference for truth' also known as the Principle of Charity (cf. Quine 1964, Davidson 2001) could lead subjects to view a sentence as true as soon as it is true on some of its readings. They would thus behave as if only the weakest reading of the sentence existed. Similar principles have been discussed by semanticists working on scope ambiguities (see Abusch 1993, Reinhart 1997, Meyer and Sauerland 2009).[5]

Now, recall that in the case of (2), the local reading asymmetrically entails the global reading, which itself asymmetrically entails the literal reading. This as such could explain why the local reading is hard to detect, even if it exists (see Sauerland 2010 for a similar

[4] In theories of type T3, the derivation of the local reading requires that the alternatives relative to which the pragmatically strengthened reading of the sentence is computed be those induced by questions of this form. In theories of type T2, the derivation of the local reading does not rely on the same alternatives. Nevertheless, the choice between various potential readings (disambiguation) is still expected to depend on what counts as relevant, and the local reading happens to be relevant typically in a context induced by a question of this form. This follows if relevance is defined in terms of answerhood, along the lines of Groenendijk and Stokhof (1984) (Danny Fox, p.c.).

[5] There are also related discussions in the psycholinguistic literature, e.g., the discussion of yes-biases by experimentalists collecting data from children. Crain and Thornton (2000) claim that young children have a yes-bias, but other works present a more complicated picture (cf. Fritzley and Lee 2003, Moriguchi, Okanda, and Itakura 2008, a.o.—thanks to a reviewer for drawing our attention to these works).

conclusion).[6]

Importantly, Geurts and Pouscoulous gave their subjects the option of reporting that a sentence could be equivocally true or false. However, the fact that the subjects do not use this option only shows that they are not aware of an ambiguity in this particular task, not that the sentence is not in principle ambiguous. Meyer and Sauerland (2009) claim that in some cases where a sentence is ambiguous between two readings R1 and R2, where R2 asymmetrically entails R1, naive subjects are not even aware of an ambiguity. Geurts and Pouscoulous showed that in other cases, involving reciprocals, subjects are able to detect an ambiguity. However, as far as we can see, the readings which underlie these judgments are not logically ordered, and so are not subject to Meyer and Sauerland's principle.[7] While there is no direct experimental evidence for Meyer and Sauerland's specific claim,[8] its plausibility is sufficient to cast doubt on Geurts and Pouscoulous' conclusions: if Meyer and Sauerland are right, the overall picture that Geurts and Pouscoulous present is consistent with the claim that the local reading exists, even though Geurts and Pouscoulous' subjects did not consciously perceive it.

[6] As pointed out by an anonymous reviewer, this reasoning seems to conflict with a principle such as the Strongest Meaning Hypothesis (SMH) (cf. Dalrymple et al. 1998), which holds that in certain cases, an ambiguity is resolved in favor of the logically strongest meaning. As Geurts and Pouscoulous (2009) correctly notes, an unrestricted version of the SMH predicts the local reading to be the preferred reading (this is arguably the case for the theory developed in Chierchia 2004). However, we believe that a version of the SMH and a principle such as 'Preference for truth' could be both simultaneously active, but at different levels. For instance, one can coherently claim a) that some principle akin to the SMH favors the generation of SIs (embedded or not) unless the resulting reading is logically weaker than the literal reading (cf. Chierchia et al. in press, Fox and Spector 2008), and b) that among the readings favored by this principle, 'Preference for Truth' makes subjects less aware of the strongest reading than they are of weaker readings. This is consistent with Meyer and Sauerland's (2009) claim that stronger readings can be detected if they are 'more accessible' than weaker readings.

In the case of sentences in which a scalar item is not embedded, there is evidence that subjects have a preference for readings with SIs (hence for the 'stronger' reading), both in Geurts and Pouscoulous' paper and in previous work (see Noveck 2001, Gualmini et al. 2001 and subsequent work). In section 5, we will present a case where the reading with an *embedded* SI is clearly preferred to the literal reading, but does *not* logically entail it. This case suggests that there is an overall preference for deriving SIs, independently of considerations of logical strength.

[7] Geurts and Pouscoulous (2009) used sentences such as 'The circles and the squares are connected with each other', and showed that they are perceived as ambiguous. They are not explicit about the nature of this ambiguity, but it seems to us that the two readings that underly the subjects' perception of an ambiguity in Geurts and Pouscoulous' particular task can be paraphrased as follows:

(i) Every circle is connected with a circle and every square is connected with a square.

(ii) Every circle is connected with a square and every square is connected with a circle.

Importantly, these two readings do not stand in an entailment relation.

[8] Meyer and Sauerland's claim is based on the following evidence: in several cases where it has been claimed in the theoretical literature, on the basis of truth-conditional intuitions, that a certain reading does not exist, the reading in question can be shown to exist by more indirect means; in all the relevant cases, the reading that had not been observed happened to be logically entailed by another reading whose existence was uncontroversial.

2.3 Summary

Geurts and Pouscoulous did not detect strong readings for sentences containing a scalar item in the scope of a universal quantifier. They interpret this result as showing that such a reading does not exist, but it appears that various factors may have conspired to make the local reading undetectable even if it existed. The experimental design that we will introduce in the next sections is designed to overcome some of the limitations of Geurts and Pouscoulous' experiments.

3 General features of our experimental design

Our own experimental design is a modified version of that of Geurts and Pouscoulous. Our aim was first to test sentences in which a scalar item occurs in the scope of a universal quantifier (experiment 1). We then tested a case where a scalar item occurs in a non-monotonic environment (experiment 2)[9]— sentences of this kind, as we will see in section 3.3, are crucial cases for comparing competing theories.

Before describing exhaustively our experimental design and our results, let us point out some of its features which we believe address the methodological limitations of Geurts and Pouscoulous' own design.

3.1 Salience, Readability and Relevance

Our pictures made it very easy to identify the items which falsify the local reading in a case where the global reading is true but the local reading is false. We used sentences such as 'Every letter is connected with some of its circles', and paired them with pictures in which the letters A, B, ..., F were surrounded by a number of circles, and possibly connected with them by a straight line. Fig. 2 represents such a picture together with the relevant sentence. This item of ours corresponds directly to the item from Geurts and Pouscoulous' experiment reported in Fig. 1: both make the global reading true and the local reading false. However, it seems to us that our pictures were easier to decipher than Geurts and Pouscoulous' pictures: in Fig. 2, identifying the letters B and E as falsifiers of the local reading is much easier than identifying the top square as a similar falsifier in Geurts and Pouscoulous' own picture (cf. Fig. 1). This addresses the point raised in section 2.2.1.

[9] An environment φ is called *monotonic* if it either preserves or reverses entailment patterns between the constituents it embeds. An environment is *non-monotonic* if it breaks entailment patterns. For instance, the quantifier 'exactly one person' creates a non-monotonic environment, because while 'eating smoked salmon' entails 'eating salmon', there is not entailment relation one way or the other between 'Exactly one person ate smoked salmon' and 'Exactly one person ate salmon'.

Every letter is connected with some of its circles.

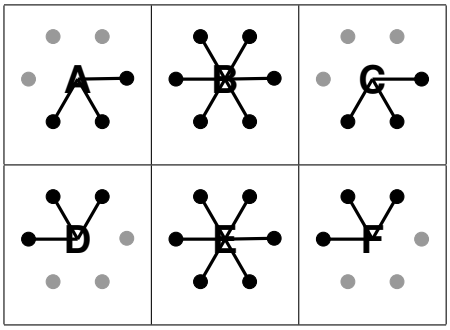


Figure 2: Example where the global reading is true, while the local reading is false. This item is a counterpart of Geurts and Pouscoulous' item represented in Fig. 1.
Note: The connected dots and the connecting segments appeared in red, and the non-connected dots appeared in gray.

We also hoped that the fact that the relevant items were different from each other (they consist of different letters from the Latin alphabet) would increase the relevance of the local reading, by drawing the subjects' attention to the individual properties of the different items, and to the way they differ from each other—thus potentially raising the following question: 'For each letter, what are its particular properties?' (cf. section 2.2.2).

3.2 Graded judgments

Instead of asking for absolute judgments of truth or falsity, we asked for graded judgments on a continuous scale ranging from 'No' (i.e. 'false') to 'Yes' (i.e. 'true')—we will provide a more detailed description of the task in section 4.1. Our expectation was that by asking for graded judgements, we would be able to bypass some of the potential consequences of the 'preference for truth' principle. By offering subjects more choices, we were likely to get more fine-grained results, which could reveal differences that remained hidden when subjects were given only two or three options.

More specifically, we made the following conjecture: given a sentence S and two distinct pictures $P1$ and $P2$, if the set of available readings for S that are true in $P1$ is a proper subset of those that are true in $P2$, then the degree to which S will be judged true will be higher in the case of $P2$ than in the case of $P1$. We thus expected that as soon as a sentence is true on one of its putative readings (relative to a certain picture), subjects would judge it true to a significant degree (due to some kind of Principle of Charity), but that this degree would increase if the sentence is evaluated with respect to a picture in which additional readings are true.[10]

[10]If this interpretation proves correct, then two kinds of hypotheses could be made to explain its cor-

In a case where the relevant sentence is true under the global reading but not under the local reading, we expected our subjects to judge the sentence to be true to a lesser degree than in a case where the sentence is true under both its local and global readings. If our interpretation of the task is correct, i.e. if the degree to which subjects judge a sentence to be true increases when more readings of the sentence are true, such a finding would provide evidence for the existence of the local reading.

The way we will interpret our results is thus based on two distinct sets of hypotheses: linguistic hypotheses about the possible readings of various sentences, and hypotheses about the experimental task itself. Our results will provide evidence for a certain combination of such hypotheses and against some others. As far as we know, our conjecture regarding the interpretation of the experimental task, namely the hypothesis that the mean rating of each condition depends on which readings of the sentence are true in this condition, has not been independently tested in the previous literature. In section 4.4.7, we will discuss a plausible alternative hypothesis regarding our experimental task, according to which the rating of a sentence-picture pair on a continuous scale reflects how close the picture is perceived to be to some prototypical situation (determined by the sentence). We will argue that even if this alternative hypothesis were correct, it would not alter our main conclusion—namely the conclusion that the local reading exists. But we will also present in section 5.5.5 additional evidence which bears on the interpretation of our experimental task and provides independent support for our initial conjecture.

3.3 Preference for truth and non-monotonic environments

In the case of sentences such as (2), the Principle of Charity, as we have seen, may make the local reading very hard to detect, because it is the logically strongest reading. However, there are cases where the computation of an SI in an embedded position is predicted to be possible by localist theories, and yet the resulting reading turns out not to be logically stronger than either the literal reading or the global reading (i.e. the reading predicted by a globalist theory). In such cases, the local reading (i.e. the reading that results from computing an SI in an embedded position), if it exists, should be easy to detect. As

rectness. One possibility is that the subjects' performances reflect the outcome of a possibly unconscious process whereby they consider all the possible readings of a sentence in parallel and determine, for each such reading, whether the sentence is true (this hypothesis is very much in line with *parallel* approaches to syntactic parsing, see e.g., Hale 2001). Another possibility (suggested to us by a reviewer) is that the subjects' mean ratings reflect the outcome of several random choices of one particular reading among the available readings, where each choice has a certain probability; for each such choice, the sentence would get a certain rating: high if the sentence is true given the chosen reading, low otherwise, but subjects might well use the scale differently from each other, and also in a non-uniform way across trials. The outcome of such a process would be similar to what is expected under the 'parallel processing' hypothesis: the mean ratings of two distinct conditions would reflect the inclusion relationships between the sets of readings that each condition makes true. Adjudicating between these two types of hypotheses is a complicated matter, orthogonal to the main goal of this paper.

we will see in section 5, sentences in which a scalar item occurs embedded under a non-monotonic operator (such as 'exactly one') provide us with precisely this kind of case.

4 Experiment 1: scalar items in universal sentences

In this experiment, we show that the local reading is available for sentences like (2) above: French scalar items like 'certains'[11] (*some*) and 'ou' (*or*), when embedded under universal quantifiers, can give rise to interpretations in which they seem to be equivalent to, respectively, 'some but not all' or an *exclusive* disjunction.

4.1 Participants and their task

16 native speakers of French ranging in age from 19 to 29 years took part in this experiment (10 women). All of them were native speakers of French and none had any prior exposure to formal linguistics.

Participants were asked to assess the truth value of a sentence in a situation which was represented graphically. The example in Fig. 3a was presented to the participants during the instructions. Crucially, they were instructed that sometimes their judgment may not be sharp and that they may thus give their answers along a continuum of answers, by positioning a cursor on a line. The example in Fig. 3b was given next to illustrate this aspect of the task. The actual instructions are reported in appendix 1. Subjects' responses were coded as a percentage of the line filled in red.[12]

4.2 Experimental items

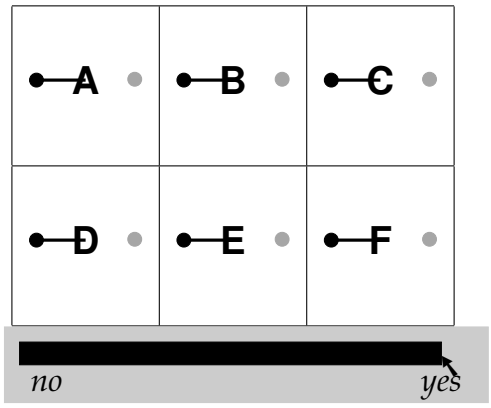
4.2.1 Target conditions: universal sentences

Each item consisted of a sentence and a picture. There were two sentences of primary interest in the experiment:

[11]Note that French *certains*, unlike its singular counterpart *un certain* or English *certain*, does not force a specific reading, especially when it is associated with a partitive phrase (*certains de*), as is the case in our experimental material. For instance, in English, the sentence 'If you solve a certain difficult problem, you will get a good grade' is necessarily interpreted as 'there is a certain difficult problem such that if you solve it, you will get a good grade'; in contrast with this, in the French sentence 'Si tu résouds certains des problèmes difficiles, tu auras une bonne note' ('*If you solve* certains *of the difficult problems, you will get a good grade*'), the phrase headed by *certains* can be interpreted in the scope of the if-clause.

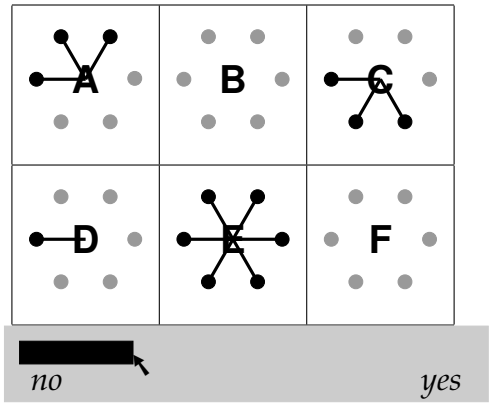
[12]This kind of measure comes directly from the magnitude estimation paradigm used in psychophysics (Stevens 1956) and already imported to linguistics to collect grammaticality judgments (see e.g., Bard, Robertson, and Sorace 1996, Schütze 1996, Cowart 1997). A similar scale of judgments was also used in Chemla (2009a,c) and Chemla and Schlenker (2009) to investigate pragmatic phenomena but not with a truth value judgment task.

Chaque lettre est reliée à son cercle rouge. *(Each letter is connected with its red circle.)*



(a) First training example in experiment 1.

Les lettres sont reliées aux cercles. *(The letters are connected with the circles.)*



(b) Second training example in experiment 1.

Figure 3: Examples presented in the instructions of experiment 1. The second example (b) was designed to illustrate that the judgments we requested were not necessarily sharp. Subjects were explicitly told that in this case it is not obvious whether the sentence is true or false, that people may disagree about it, and that they were requested to position the cursor to represent their intuition. For more details, see the actual instructions in appendix 1.
Note: In every 2-dot configuration, the left circle (and the segment connecting it to the letter, if any) was red and the right one was blue. In every 6-dot configuration, the connected circles, as well as the connections themselves, were red, while the other dots were gray. The response showed as a red line of variable length on a gray background.

(8) Chaque lettre est reliée à certains de ses cercles.
Each letter is connected with some of its circles.

(9) Chaque lettre est reliée à son cercle rouge ou à son cercle bleu.[13]
Each letter is connected with its red circle or with its blue circle.

We are interested in the following potential readings of these sentences (cf. (7)):

(10) Possible readings of (8):
   a. Literal Reading: Each letter is connected with at least one of its circles.
   b. Global Reading: Each letter is connected with at least one of its circles, and it is not the case that each letter is connected with all its circles.
   c. Local Reading: Each letter is connected with at least one of its circles, and no letter is connected with all its circles.

(11) Possible readings of (9):
   a. Literal Reading: Each letter is connected with at least one of its two circles.
   b. Global Reading: Each letter is connected with at least its blue circle or its red circle, and it is not the case that each letter is connected with both its circles.
   c. Local Reading: Each letter is connected with either the blue circle or the red circle, and no letter is connected to both.

Each of these sentences was paired with various pictures, giving rise to the following four target conditions (see illustrative examples in Fig. 4): **FALSE:** no reading is true, **LITERAL:** only the literal reading is true, **WEAK:** both the literal and the global readings are true but the local reading is false, **STRONG:** all readings are true.[14]

Representative examples of pictures corresponding to each of these conditions are given in Fig. 4. The entire set of pictures used to instantiate these conditions in the experiment is described in appendix 2.1.

4.2.2 Downward entailing (DE) environments

Both localist and globalist theoreticians agree that the embedded SIs in downward-entailing environments are, at best, marginal.[15]

[13]The colors and color names changed according to the picture. Color names were always monosyllabic: 'rouge'(*red*), 'bleu'(*blue*) or 'vert'(*green*). No participant was color blind.

[14]These conditions represent all possible combinations of true and false readings because of the entailment relations between the readings: in both (10) and (11), the local reading (c) entails the global reading (b) which entails the literal reading (a). Hence, whenever the local reading is true, the two others are automatically true as well, and whenever the global reading is true, the literal reading is true as well.

[15] An environment φ is *downward-entailing* if it licenses inferences from supersets (e.g., 'salmon') to subsets (e.g., 'smoked salmon'): 'φ(salmon)' entails 'φ(smoked salmon)' . This can be seen as a generalized notion of negativity: 'John didn't eat *salmon*' entails 'John didn't eat *smoked salmon*'. Upward-entailingness (UEness) is the reversed notion (the inference should be in the other direction, as in 'John ate *smoked salmon*' entails 'John ate *salmon*') and non-monotonicity describes environments that are neither DE nor UE.

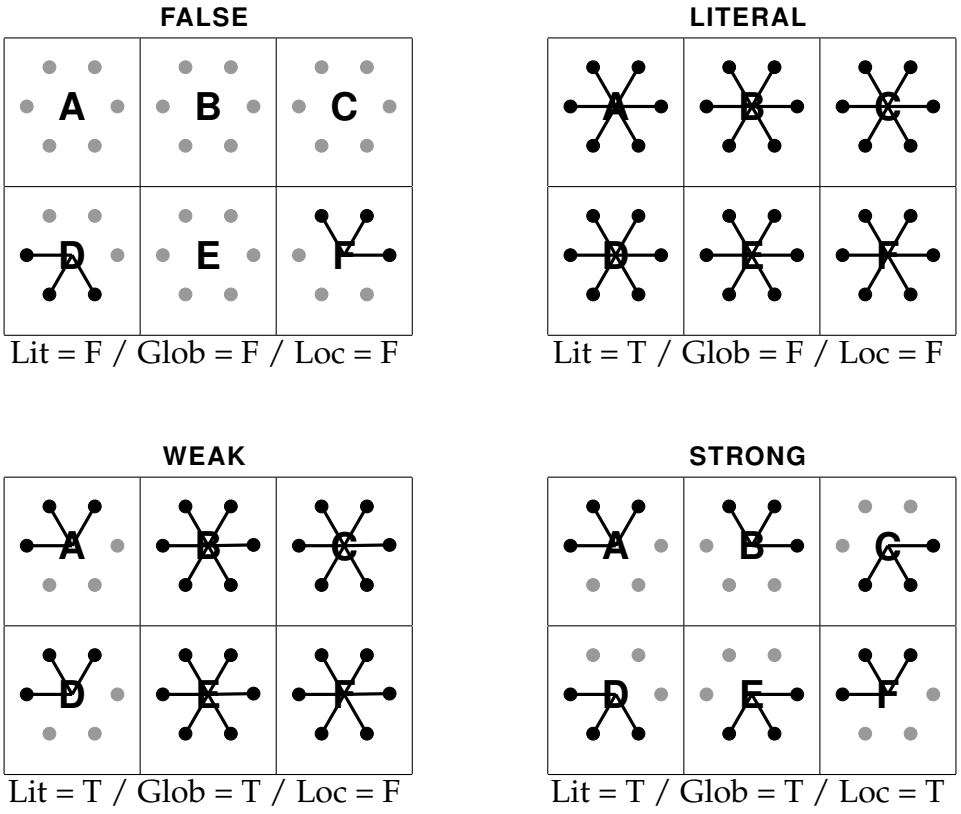


Figure 4: Illustrative examples of the images used to illustrate the different conditions **FALSE**, **LITERAL**, **WEAK** and **STRONG** for the test sentence (8): 'Every letter is connected with some of its circles'. We also reported below each image whether the literal (Lit), global (Glob) and local (Loc) readings are true (T) or false (F).

For instance, when scalar items are embedded in the scope of 'No' as in (12) or (13), it is uncontroversial that the potential 'local' readings described in (14) and (15) do not normally arise, unless a particular intonation is used:[16]

(12) Aucune lettre n'est reliée à certains de ses cercles.
No letter is connected with some of its circles.

[16]Within the neo-Gricean approach, it is predicted that scalar items cannot retain their 'strong' reading in downward-entailing environments; however, it has long been recognized that, with a particular intonation pattern, such 'local' readings are possible (see, e.g., Levinson 2000 for a discussion of such 'intrusive implicatures'). Some neo-Gricean authors accommodate this fact by introducing a special mechanism, such as a 'metalinguistic' mechanism (see, e.g., Horn 1985, 2006) or a 'reconstrual' mechanism (Geurts 2009). Localist approaches impose a constraint that makes the relevant reading either impossible (Chierchia 2004) or dispreferred (Chierchia et al. in press). See also footnote 6.

(13) Aucune lettre n'est reliée à son cercle rouge ou à son cercle bleu.
No letter is connected with its red circle or with its blue circle.

(14) Potential local reading of (12): No letter is connected with some but not all of its circles.

(15) Potential local reading of (13): No letter is connected with exactly one of its two circles.

These examples provide an interesting point of comparison for our purposes: unless participants are influenced by some artificial experimental strategy, they should either fail to perceive the 'local' reading for such sentences, or they should perceive it as only marginally available (given the marginal availability of the local reading, cf. footnote 16). Hence, we included (12) and (13) at the end of the experiment, paired with pictures leading to the following three types of conditions: **FALSE:** no reading is true (e.g., in the case of 'some', each letter was then connected with a strict subset of its circles), **?LOCAL:** only the local reading is true (e.g., each letter was connected with all its circles), **BOTH:** both the local and the literal readings are true (e.g., each letter was connected with none of its circles). A more systematic description of the pictures relevant for each condition is given in Appendix 2.2.

4.2.3 Procedure

The instructions and the training items were presented first to allow participants to get used to the display and to the task (see Fig. 3a and 3b). After that, participants were given two blocks of test items with a short break in between. All target conditions appeared several times in each block. More specifically, for each scalar item, the **FALSE** condition was presented with 6 different items, **LITERAL** with 2, **WEAK** with 8, **STRONG** with 4 items (see Table 1 for more details). The DE conditions were administered in a third block which contained no target items.[17] In each block, the items were presented in pseudo-random order.

4.3 Predictions

The theories we described differ as to whether they predict that the local readings of (8) and (9) exist or not. Theory T1 predicts that the local reading does not exist and thus anticipates no difference between the conditions **WEAK** and **STRONG**. On the other hand, theories T2 and T3 predict that the local reading exists, and therefore that the relevant sentences will receive a higher score (measured by the mean position of the cursor) in

[17] There was no break before the last block but we varied the colors of the circles and the sentences were written in bold to make sure that the subjects would notice that the sentences were not the same as before.

the **STRONG** condition than in the **WEAK** condition—since one more reading is true in the **STRONG** condition.

4.4 Results and interpretation

In this section, we present the results of the experiment and draw a number of conclusions. The main result is given in section 4.4.2 and two control results are described in section 4.4.3 and section 4.4.4. In the following subsections, we report two post-hoc analyses, one about 'distributivity inferences' (section 4.4.5), and another one which relates to a plausible alternative hypothesis about what graded truth-value judgments measure (section 4.4.6).

4.4.1 Preliminary technical remarks

We lost 5.0% of the responses in target conditions for technical reasons.[18] All statistical analyses of pairwise differences reported below are computed per subject ($n = 16$) using Wilcoxon rank-sign tests (the W values reported correspond to the maximum of W_- and W_+). Mann-Whitney U tests were also computed by item, except in the DE conditions, in which there were too few items to support such an analysis. The items analyses, where conducted, yielded similar results to the subjects analyses. Statistical analyses of interactions were computed using standard parametric analyses of variance (ANOVA) for simplicity.[19]

Finally, the statistical analyses are reported without correction for multiple comparisons, but the p values are reported with different significance levels ($< .05$, $< .01$, $< .005$ etc.). Hence, one can immediately check that all key results remain statistically significant at the $.05$ level when a Bonferroni correction *à la* Holm for 5 simultaneous comparisons is applied (practically, given the level of significance that we report, this amounts to checking that in a family of 5 or less, all comparisons are significant at the level $p < .05/5 = .01$, except possibly for one which may be significant at $p < .05$).

4.4.2 Main result: detection of the local reading

Fig. 5 reports the mean ratings in the target conditions. All differences between two consecutive bars are significant as revealed by the corresponding Wilcoxon signed-rank tests ($n = 16$) both for the item 'some' (**FALSE** vs. **LITERAL**: $W = 123$, $p < .005$; **LITERAL** vs. **WEAK**: $W = 134$, $p < .001$; **WEAK** vs. **STRONG**: $W = 134$, $p < .001$) and 'or' (**FALSE** vs.

[18] There were mainly two primary reasons for data loss: 1) failure to register a response, possibly because the response was given before the item was fully loaded and 2) because the data were automatically sent to an internet database, internet connection problems occasionally resulted in lost responses.

[19] Most statistical analyses presented in the paper, including two-way interactions, were also computed with bootstrap procedures. They always led to the same conclusions.

LITERAL: $W = 108$, $p < .05$; **LITERAL** vs. **WEAK**: $W = 129$, $p < .001$; **WEAK** vs. **STRONG**: $W = 133$, $p < .001$).

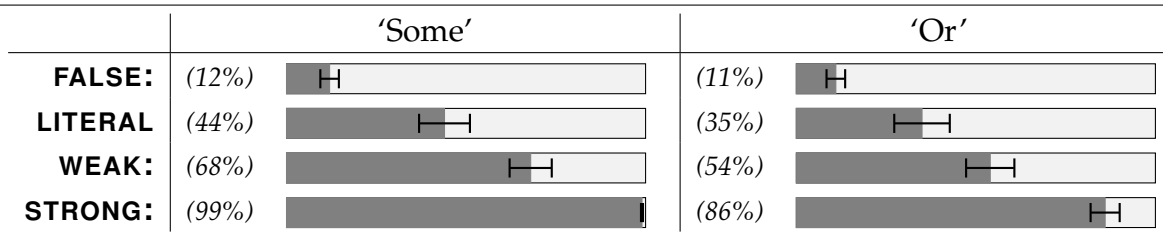


Figure 5: Main results: Mean position of the cursor in the target conditions of Exp. 1 (see section 4.2.1 or Fig. 4 for an illustration). Error bars represent standard errors to the mean.

The crucial part of this result is that the ratings are higher in the **STRONG** condition than in the **WEAK** condition, even though the two conditions differ only according to the truth value of the local reading. This difference provides important support for the existence of the local reading. Indeed, these results are fully explained if we assume that a) the target sentence is ambiguous between the literal reading, the global reading and the local reading, and b) the more readings are true, the higher the sentence is rated. They are not expected if only the literal and the global readings exist.

4.4.3 Analyses of changes in performance between the two experimental blocks

As discussed above (section 4.2.3), the items were split in two consecutive, formally identical blocks. This partially allows us to check whether the effects we detected should be attributed to some response strategy developed in the course of the experiment. There is no reliable ground for saying so from our data since the global 2(Block) $\times$ 4(Condition) ANOVA does not reveal a significant interaction ($F(3, 45) = 1.9$, $p = .15$). The main effect of Condition is significant ($F(3, 45) = 86$, $p < .001$) and the main effect of Block is not ($F(1, 15) = 3.0$, $p = .11$).

Similar analyses restricted to each scalar item yield similar results: no significant interaction between Block and Condition ('some': $F(3, 45) = 2.2$, $p = .10$, 'or':[20] $F(3, 45) = .95$, $p = .43$), a main effect of Condition ('some': $F(3, 45) = 67$, $p < .001$, 'or': $F(3, 45) = 38$, $p < .001$), and no main effect of Block ('some': $F(1, 15) = 3.1$, $p = .099$, 'or': $F(1, 15) = 1.8$, $p = .20$).

[20]In this particular case, the overall loss of data resulted in one empty cell (missing value): for one participant, the **LITERAL** condition was missing in the first block of items. The value reported in the text was computed by replacing the one missing value by the mean of values for the other subjects in the same condition. The statistical results remain the same without replacing the missing value ($F(3, 44) = .81$, $p = .50$) or excluding the participant with incomplete data ($F(3, 42) = .82$, $p = .50$). The statistical results also remain the same with these modifications for the main effects of Block and Condition.

These analyses do not reveal a *significant* change of results through time. Yet we cannot formally exclude that subjects were influenced, for instance, by repeated exposure to our pictures(see sections 5.5.3 and 5.5.4 for related analyses and discussion). In any case, even if there was such an effect, this would not alter our conclusion that speakers can access local readings in some contexts.

4.4.4 Analyses of responses for downward-entailing environments

Fig. 6 reports the results for the DE conditions described in section 4.2.2. For the scalar item 'some', the relevant tests show a significant difference within all pairs of conditions (**FALSE** vs. **?LOCAL**:[21] $W = 65.5$, $p < .05$; **?LOCAL** vs. **BOTH**: $W = 130$, $p < .005$). For the scalar item 'or', there is no difference between the **FALSE** condition where the sentence is unambiguously false and the **?LOCAL** condition, where the sentence could be judged true because of the local reading (**FALSE** vs. **?LOCAL**:[22] $W = 47$, $p = .56$, **?LOCAL** vs. **BOTH**: $W = 133$, $p < .001$).

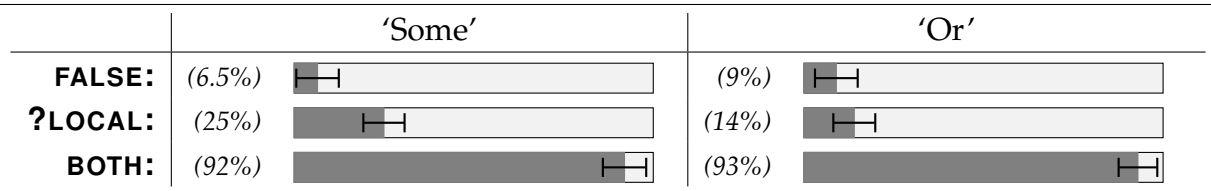


Figure 6: Mean responses for the DE control conditions in exp. 1 (see section 4.2.2).

The fact that the **?LOCAL** condition is judged a little higher than the **FALSE** for the item 'some' suggests that the local reading is perceived to a certain extent, contrary to what we expected. This is nevertheless not terribly disturbing, for two reasons. First, it does not generalize to the scalar item 'or'. Second, what is important for us is that the control sentences receive a low rating in the condition **?LOCAL**, compared, e.g., to ratings of conditions where it is uncontroversial that the target sentence has a true reading.[23] The small difference we found in the case of 'some' might reflect the fact that, as mentioned in footnote 16, the 'local' reading, though normally strongly dispreferred under the scope of negation, is nevertheless thought to be marginally available even in these cases.

[21] p is computed with $n = 12$ because of ties.

[22] p is computed with $n = 12$ because of ties.

[23] Note that even with the scalar item 'some', the condition **?LOCAL** is rated at a radically lower level than the condition **BOTH** (25 % vs. 92 %); more importantly, in the case of 'some', the condition **?LOCAL** is rated much lower than conditions in which it is uncontroversial that the target sentence has a true reading (for instance, the difference between this **?LOCAL** condition and the **WEAK** condition from the main target sentences is statistically significant: $W = 129$, $p < .001$).

Overall, the effect of the potential local reading seems to be clearly different in universal contexts and in DE contexts.

4.4.5 Analyses of 'distributivity' effects

Different pictures instantiated each target condition, and some differences between these pictures could be of importance. In Appendix 2.1 we give a complete description of the pictures we used, and Appendix 3.1 reports fine-grained results, where different instantiations of each condition are not aggregated. In this section and the next one, we present two aspects of these fine-grained, post-hoc analyses.

So far, we have disregarded one aspect of the target sentence (9) repeated below as (16), namely the fact that this sentence triggers the following 'distributivity' inferences:

(16) Every letter is connected with its red circle or with its blue circle.

$\rightsquigarrow$ Some letters are connected with their red circle and not with their blue circle.

$\rightsquigarrow$ Some letters are connected with their blue circle and not with their red circle.

These inferences are particularly relevant for our purposes because various theories suggest that they have the same theoretical status as the scalar implicatures we are interested in (cf. Klinedinst 2006, Spector 2006, Fox 2007, Chemla 2008, 2009b, very much in the spirit of work comparing SIs with free choice effects as in e.g., Kratzer and Shimoyama 2002, Schulz 2003). Interestingly, the experimental items instantiating the **STRONG** condition included cases where these inferences are supported (**STRONG[$\neq$]**) and cases where they are not (**STRONG[$=$]**).[24] These configurations are illustrated in Fig. 7.

Crucially, participants judged higher **STRONG[$\neq$]** cases where these distributivity inferences are supported (99.5%) than **STRONG[$=$]** cases where they are not (73%): $W = 78$, $p < .005$.[25] This new result fits well with the previous results: the sentence received its highest score exactly when *all* the inferences it can normally give rise to are supported.

4.4.6 Does the exact number of verifiers matter?

In section 4.2.1, we distinguished the conditions according to which a reading was true or false. The pictures we used in the experiment can be distinguished more finely according to *how many* letters contribute to make each reading true or false. Hence, the **FALSE** condition can be split in three sub-conditions **FALSE-0**, **FALSE-2** and **FALSE-4** depending on whether 0, 2 or 4 letters actually satisfy the predicate (see Fig. 8) .

[24] We use the $=/\neq$ notation because the distributivity inferences correspond to *differences* between the items: if the sentence and its distributivity inferences are true, some letter should be connected only with its red circle and some other letter should be connected only with its blue circle.

[25] As can be seen in Fig. 15, no similar effect is found with (8) "Each letter is connected with some of its circles" (answers: 99% and 98%). This absence of difference is not surprising since 'some' does not give rise

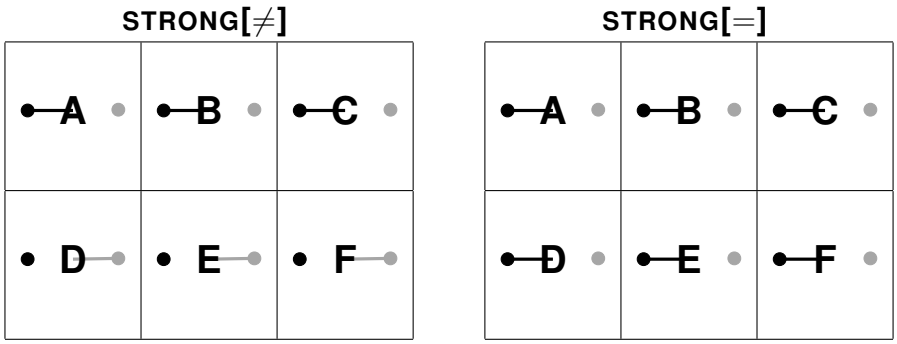


Figure 7: Two instantiations of the **STRONG** conditions for the test sentence (9). The distributivity inferences are supported in **STRONG[≠]**, but they are not in **STRONG[=]** (because every letter is connected with its red circle). See note on colors in Fig. 3b.

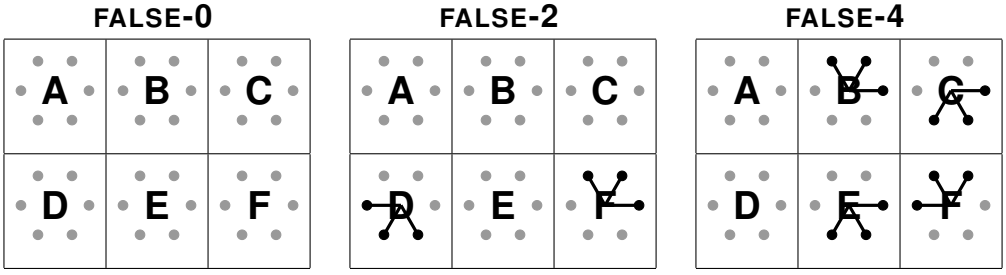


Figure 8: Illustrative examples of sub-conditions of the **FALSE** condition for the 'some' sentence: no letter satisfies the predicate (**FALSE-0**), 2 letters satisfy the predicate (**FALSE-2**), or 4 letters satisfy the predicate (**FALSE-4**).

As illustrated in Fig. 9, the intermediate condition **WEAK** can be split in **WEAK-2** and **WEAK-4** depending on whether 2 or 4 letters make the *strong* version of the predicate true, i.e. whether 2 or 4 letters are connected with some *but not all* of their circles.

Figure 10 represents the results split according to these new sub-conditions. Focussing on the false items first, notice that they are not judged equally false.[26] This shows that participants' answers are influenced by the number of items satisfying the predicate: if there are more items satisfying P, participants rate the sentence 'Each x $P(x)$' higher (even though the sentence remains false as long as not all xs satisfy P).

Similarly, when the literal meaning of the sentence is true, the more *strong* verifiers

to comparable distributivity inferences.

[26]The Friedman rank test (testing non-parametrically the null hypothesis that all **FALSE** sub-conditions are judged the same) yields a significant outcome both for 'some' ($\chi^2(2) = 20$, $p < .001$) and for 'or' ($\chi^2(2) = 17$, $p < .001$). The relevant pairwise differences are significant (all but one at the level $p < .005$).

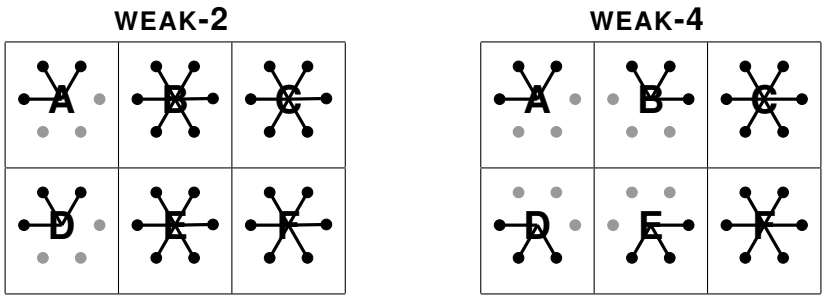


Figure 9: Illustrative examples of the sub-conditions of the **WEAK** condition for the 'some' sentence: 2 letters satisfy the strengthened predicate (**WEAK-2**) or 4 letters satisfy the strengthened predicate (**WEAK-4**).

there are, the higher the sentence is rated. In particular, the successive differences between **WEAK-2**, **WEAK-4** and **STRONG** are all significant.[27]

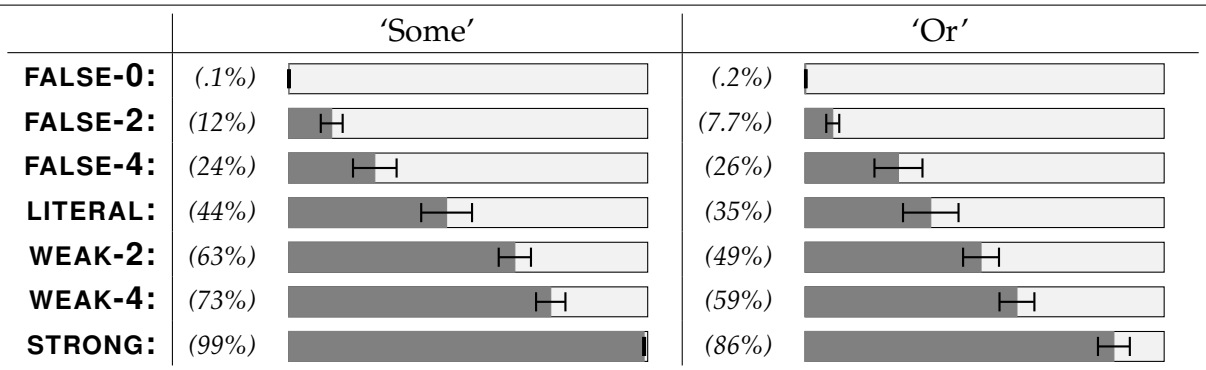


Figure 10: Detailed results including sub-conditions where a different number of letters contribute to make various readings true or false.

4.4.7 An alternative interpretation: graded judgements as typicality judgments

On the face of these results, the following hypothesis might seem plausible: the rating of each condition is driven mainly by the number of strong verifiers which occur in the picture, rather than by the truth-values of the three specific readings that we have hypothesized. For instance, the reason why the **STRONG** condition receives the highest ratings is that it contains as many strong verifiers as possible. This suggests the following

[27] Statistics for 'some': **WEAK-2** vs. **WEAK-4**: $W = 114$, $p < .005$, **WEAK-4** vs. **STRONG**: $W = 131$, $p < .001$. For 'or': **WEAK-2** vs. **WEAK-4**: $W = 112$, $p < .05$, **WEAK-4** vs. **STRONG**: $W = 128$, $p < .001$.

alternative interpretation of the subjects' behavior (let us call it the 'typicality interpretation'): when asked to use a graded scale, subjects tend to rate the relevant sentences not simply on the basis of their perceived truth-conditions, but rather in terms of some *metric* which reflects the 'distance' between a particular situation and some 'prototypical situation' which is determined by the relevant sentence.[28] According to the typicality interpretation, the differences between the ratings of different conditions do not reflect (or at least do not *only* reflect) an inclusion relationship between the sets of readings that each condition makes true, and there would be no need to assume that the relevant sentence is ambiguous at all. The fact that the **FALSE-4** condition receives such a high rating compared to the **FALSE-0** condition (24% vs. 0.1%) suggests that typicality is one of the relevant factors explaining our results.

However, one should ask the following question: if the main factor explaining these results is the one hypothesized by the 'typicality interpretation', what must the underlying metric be? More specifically, what kind of situations must be counted as 'prototypical' instances of the sentence? As far as we can see, one should conclude that the best instances of the sentence among our various pictures are the ones used in the condition that receive the highest rating, namely the condition **STRONG**. Note that there is a 10% difference between the sub-conditions of the **WEAK** conditions (i.e. between **WEAK-2** and **WEAK-4**), while the difference between these cases and the **STRONG** condition is twice as big (even more so if we factor out distributivity inferences for 'or', see section 4.4.5). We would thus be led to conclude that the best instances of the sentence are those which make the local reading true. But it is hard to see how this could be so if the local reading did not correspond to a salient reading of the sentence, or at least if the inferences that correspond to the local reading were not strongly supported by the sentence.[29] So the 'typicality interpretation', as far as we can see, would support our conclusion that the

[28] Armstrong, Gleitman, and Gleitman (1983) present evidence that the use of a graded scale prompts subjects to assess the degree to which a given object is a 'typical' instance of a concept, even in cases where there is not doubt that the object in question *is* an instance of the concept: for instance 4 is judged to be a better instance of an even number than 5172, even though, when asked explicitly, the same subjects stated that 'even number' is not a graded category.

[29] In principle, one could imagine that the underlying metric might be defined not only in terms of the closeness of a given picture to some typical situation that makes the sentence true, but also in terms of its *remoteness* from some typical situation that makes the sentence *false*. One could then try to argue that the only salient reading of the sentence is the global reading, as Geurts and Pouscoulous claim, and that the observed pattern is to be explained as follows: the reason why the **STRONG** condition is rated the highest is that, *among the conditions that makes the literal reading true*, it is the one that is the most *remote* from a case that makes the global reading *false*. However, this interpretation seems to us not to be plausible, because it is based on a metric which completely ignores the situations corresponding to the **FALSE** condition, in which both the literal and the global readings are false: on any reasonable metric, the **STRONG** condition is *closer* to these cases of falsity than the **WEAK** and **LITERAL** conditions are, so that it is absolutely unclear why the **STRONG** condition should count as more 'typical' if one hypotheses a metric based on remoteness from cases of falsity.

local reading exists.

Note also that the typicality interpretation and our initial hypothesis are not mutually exclusive. The fact that typicality seems to play a role does not as such invalidate our initial interpretation. The ratings of each condition could be a function both of the set of readings that are true in this condition and of the 'closeness' of the picture to the 'typical' instances of each reading. A more complex interpretation of this sort would still support, as far as we can see, the claim that the local reading is a possible reading for the sentence.

As we will see in subsections 5.5.5 and 5.5.6, the results of our second experiment actually provide support for our initial hypothesis, which relates the mean rating of a condition to the set of readings that are true in this condition (which is not to say that typicality considerations do not also play a role in subjects' behavior). First, in these cases, there will simply be no plausible alternative interpretation of the data solely in terms of typicality. Second, we will be able to extract from our results independent evidence that graded judgments are able to reveal certain ambiguities.

4.5 Experiment 1: summary

The main result of this experiment is that participants rate the target sentences higher when both the local and the global readings are true than when the global reading is true but the local reading is false (see the difference between **WEAK** and **STRONG** reported in section 4.4.2). This provides a strong argument that the local reading exists, contra Geurts and Pouscoulous' premature conclusions. However, as discussed earlier, the existence of the local reading is not sufficient to distinguish between globalist and localist theories of scalar implicatures. In fact, several current globalist theories manage to generate this reading.

Our second experiment is based on the same methodology. But by testing non-monotonic environments, it will tackle more directly the debate between globalist and localist approaches to scalar implicatures.

5 Experiment 2: scalar items in non-monotonic environments

5.1 Background

In this new experiment, we tested cases for which pragmatic and grammatical theories are bound to make different predictions. This happens with sentences where a scalar item like 'some' or 'or' occurs in a non-monotonic environment:[30]

(17) Exactly one letter is connected with some of its circles.

(18) Exactly one letter is connected with its blue circle or with its red circle.

[30] See footnote 15 for definitions of non-monotonic, and downward- and upward-entailing environments.

The relevant potential readings (i.e. those that the sentence could in principle have according to various theories) can be paraphrased as follows:

(19) Potential readings of (17)

a. Literal meaning: one letter is connected with some or all of its circles, the other letters are connected with no circle.

b. Global reading: one letter is connected with some but not all of its circles, the other letters are connected with no circle.

c. Local reading: one letter is connected with some but not all of its circles, the other letters may be connected with either none or all of their circles.

(20) Potential readings of (18)

a. Literal meaning: one letter is connected with its blue circle or with its red circle or with both, the other letters are connected with no circle.

b. Global reading: one letter is connected with exactly one of its two circles, the other letters are connected with no circle.

c. Local reading: one letter is connected with exactly one of its two circles, the other letters may be connected with either none or both of their circles.

The global reading (19b) is obtained by adding to the literal reading the negation of the alternative sentence "Exactly one letter is connected with all its circles".[31] On the other hand, the local reading (19c) is obtained by interpreting 'some' as equivalent to 'some but not all'.

Importantly, because the scalar item now occurs in a non-monotonic environment, the local reading does not entail the global reading. In fact, it does not even entail the literal reading. Rather, the logical relationships between the three potential readings are as follows: the global reading entails both the literal and the local reading, and the literal reading and the local reading are logically independent of each other. This is of major importance for three reasons.

First, globalist theories are bound to predict readings that entail the literal reading: they predict that the negation of certain alternatives is *conjoined* with the literal reading. Hence, even if some Gricean theories can derive a seemingly local reading for universal sentences, they cannot predict local readings like (19c) or (20c) in these non-monotonic cases. Second, the fact that the local reading does not entail any of the other two potential readings could automatically make it easier to detect, see section 2.2.3. Finally, this very fact will also allow us to construct cases where only the local reading is true and to assess its existence independently of the other readings.

[31] We let the reader check that the conjunction of (17) and the negation of "Exactly one letter is connected with all its circles" is equivalent to (19b). Notice that the alternative negated to derive the global reading is not stronger than the original sentence; cf. footnote 1 and more specifically, for a discussion of non-monotonic contexts, Spector (2007a), Chemla (2008, 2009b) and Chierchia et al. (in press).

5.2 Participants and their task

16 native speakers of French ranging in age from 18 to 35 years took part in this experiment (9 women). All of them were native speakers of French and none had any prior exposure to formal linguistics.

The task was the same as for experiment 1. The instructions were also identical except that, because the pictures for the target items involve 3-grids for experiment 2 instead of 6-grids for experiment 1, the training items were modified accordingly.

5.3 Experimental items

5.3.1 Target conditions

Contrary to experiment 1, the target sentences were not universal sentences but involved the non-monotonic quantifier 'exactly 1':

(21) Il y a exactement une lettre reliée à certains de ses cercles.
There is exactly one letter connected with some of its circles.

(22) Il y a exactement une lettre reliée à son cercle rouge ou à son cercle bleu.
There is exactly one letter connected with its red circle or with its blue circle.

Each of these sentences was paired with various pictures, giving rise to the following four target conditions: **FALSE:** no reading is true, **LITERAL:** only the literal reading is true, **LOCAL:** only the local reading is true and **ALL:** all three readings—literal, global and local—are true.[32] Representative examples of pictures instantiating each of these conditions are given in Fig. 11. The whole list of pictures used in the experiment is described in appendix 2.3.

5.3.2 Downward entailing environments

The same conditions we used in experiment 1 were included in this experiment as well, for identical reasons (see details in section 4.2.2). Notice that these items now differ from the rest of the experimental items in that they are constructed from 6-grids as before, while the rest of the experimental items are now 3-grids.

5.3.3 Presentation of the items

The items were presented just like in experiment 1 (see section 4.2.3): the examples from the instructions were presented first, then came two blocks of target conditions, and finally came a block with the DE control conditions.

[32] These conditions represent the whole range of combination of true and false readings because of their entailment relations.

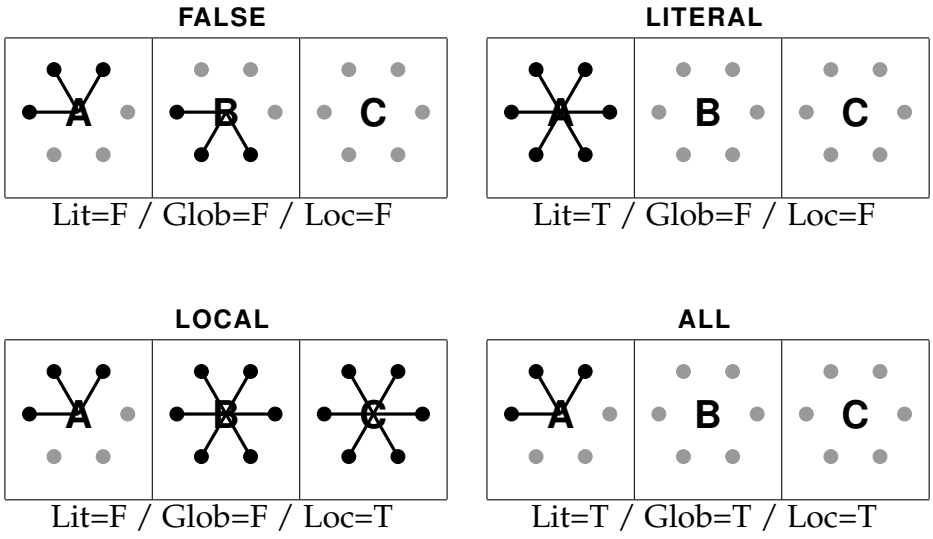


Figure 11: Illustrative examples of the images used to illustrate the different conditions **FALSE**, **LITERAL**, **LOCAL** and **ALL** for the test sentence (21): 'Exactly one letter is connected with some of its circles'. We also reported below each image whether the literal (Lit), global (Glob) and local (Loc) readings are true (T) or false (F).

5.4 Predictions

In non-monotonic contexts, localist theories predict that the local reading exists while globalist theories cannot derive this reading. Moreover, in the **LOCAL** condition, the local reading is true, while all the readings predicted by globalist theories are false. Hence, in the **LOCAL** condition (see Fig. 11 for an example), globalist theories predict that the sentence is plainly false, while the localist theories predict that the sentence has a true reading.

5.5 Results and interpretation

5.5.1 Preliminary technical remarks

We lost 15% of the responses in target conditions for technical reasons (see footnote 18). See section 4.4.1 for more details about the reported statistical analyses.

5.5.2 Main result: the local reading exists

Fig. 12 reports the mean ratings of the target items grouped according to which interpretation is true: none, local only, literal only, all. All pairwise differences are significant,

except for the **LOCAL** vs. **LITERAL** conditions in the case of 'or'.[33] (The relevant Wilcoxon tests for 'some': **FALSE** vs. **LITERAL**: $W = 126$, $p < .005$, **LITERAL** vs. **LOCAL**: $W = 109$, $p < .05$, **LOCAL** vs. **ALL**:[34] $W = 105$, $p < .005$; and for 'or': **FALSE** vs. **LITERAL**: $W = 123$, $p < .005$, **LITERAL** vs. **LOCAL**: $W = 92$, $p = .23$, **LOCAL** vs. **ALL**:[35] $W = 120$, $p < .001$).

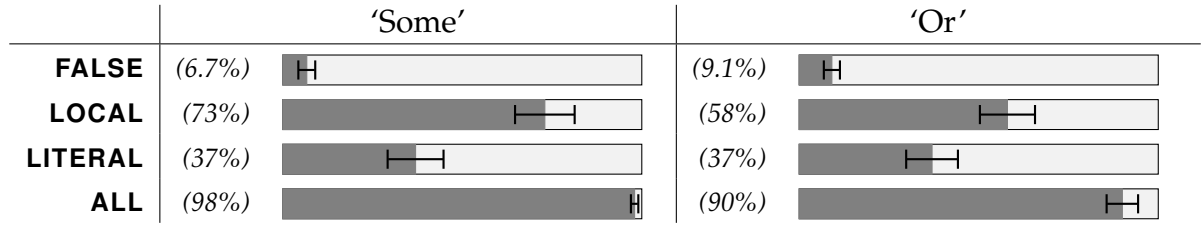


Figure 12: Mean responses in the target conditions of experiment 2 (see section 5.3.1 or Fig. 11 for an illustration).

This first set of data qualifies the local reading as a possible interpretation of non-monotonic sentences since the **LOCAL** condition is rated much higher than in the **FALSE** condition, and is in fact rated very high (73% for the sentence with 'some' and 58% for the sentence with 'or').

Furthermore, the **LOCAL** condition is rated significantly higher than the **LITERAL** condition, a fact which is unexpected under the globalist approach, but can be understood within the localist approach. Specifically, this fact suggests that the preference for readings which include SIs (over readings without any SIs), noted in the literature, is not specifically a preference for *global* SIs, but rather a general preference for deriving SIs, be they embedded or not-embedded (unless the resulting reading is weaker than the literal reading, as is the case when an SI is embedded in a DE-environment). Note also that this preference cannot be explained by a principle like the Strongest Meaning Hypothesis, since it is observed even in this case, where the resulting SI reading is not stronger than the literal reading (cf. footnote 6).[36]

5.5.3 Analyses of changes in performance between the two experimental blocks

The items were presented in two consecutive similar blocks. Yet, the 2(Block) $\times$ 4(Condition) ANOVA shows no significant interaction ($F(3, 45) = 1.2$, $p = .31$). The same ANOVA reveals a significant main effect of Condition ($F(3, 45) = 41$, $p < .001$) and no robust main effect of Block ($F(1, 15) = 1.2$, $p = .29$).

[33] On a per item analysis, this difference does come out significant : $U = 32$ ($n_1 = 4, n_2 = 8$), $p < .005$.
[34] p is computed with $n = 14$ because of ties.
[35] p is computed with $n = 15$ because of one tie.
[36] Thanks to an anonymous reviewer for making this point.

Similar analyses restricted to each item yield similar results: no reliable interaction between Block and Condition ('some': $F(3, 45) = 2.2$, $p = .11$, 'or': $F(3, 45) = .14$, $p = .93$), a main effect of Condition ('some': $F(3, 45) = 45$, $p < .001$, 'or': $F(3, 45) = 30$, $p < .001$), and no reliable main effect of Block ('some': $F(1, 15) = 2.4$, $p = .14$, 'or': $F(1, 15) = .21$, $p = .65$).[37]

Thus we did not find any *significant* change of results across time, as in experiment 1. As we have already pointed out, this does not exclude that subjects were influenced by repeated exposure to our pictures (see section 4.4.3). The results we discuss in the next section (DE-conditions) actually suggest that repeated exposure to our conditions may modify the behavior of subjects across time, in certain cases at least.

5.5.4 Analyses of responses for downward entailing environments

Fig. 13 reports the results for the control DE conditions described in section 5.3.2. Statistical tests reveal significant differences for both pairs of conditions with 'some' (**FALSE** vs. **?LOCAL**: $W = 112$, $p < .005$; **?LOCAL** vs. **BOTH**:[38] $W = 91$, $p < .005$), and only for the **?LOCAL** vs. **LITERAL** comparison with 'or' (**FALSE** vs. **?LOCAL**: $W = 105$, $p = .058$; **?LOCAL** vs. **LITERAL**: $W = 136$, $p < .001$).

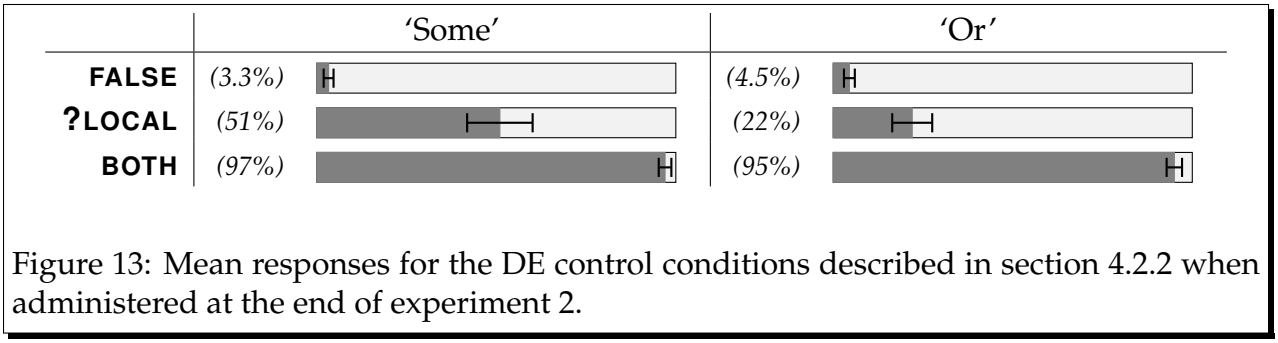


Figure 13: Mean responses for the DE control conditions described in section 4.2.2 when administered at the end of experiment 2.

Surprisingly, the rates for the **?LOCAL** condition are higher than they were in the first experiment (compare Fig. 13 to Fig. 6).[39] Importantly, though, the local reading in DE contexts appears to be less accessible than the derivation of local implicatures in non-monotonic environments from the main part of the experiment ('some': 51% vs. 73%: $W = 98$, $p < .05$; 'or': 22% vs. 58%: $W = 108$, $p < .01$).[40]

[37] The overall loss of data resulted in four missing values: for one participant, the **LOCAL** condition was missing for the item 'or' in the first block and in the second block for the item 'some'; for two other participants the **LITERAL** condition was missing for the item 'some' in the second block. The values reported in the text were computed by replacing the missing values by the mean of the corresponding values for the other subjects in the same condition. We checked that the same result would be obtained without these replacements. See more details in section 4.4.3, footnote 20.

[38] The p-value was computed with $n = 13$ because of two ties and one missing value.

[39] Notice however that this comparison involves different groups of participants.

[40] The p-value was computed with $n = 15$ because of one tie.

The very same control conditions thus received very different scores depending on whether they were presented at the end of the first experiment or at the end of the second experiment. So far, we do not have a clear understanding of this difference, but we would like to suggest an explanation along the following lines. As we briefly mentioned in section 4.2.2 (cf. in particular footnote 16), embedded SIs in downward-entailing environments are thought to be strongly dispreferred, but not absolutely impossible. It could be that subjects become much better at perceiving 'local' readings even in cases where they are normally disprefered *once* they have experienced cases in which the local reading is salient. The target conditions of the second experiment seem to have precisely this property. We anticipated that the local reading would be easier to identify as a separate reading in the second experiment than in the first experiment because it was logically independent of the literal reading, and not stronger than the global reading. The results we have just presented for these target sentences of the second experiment confirmed that subjects were very good at perceiving the local reading.

5.5.5 More fine-grained results: detection of a scope ambiguity

It is known that a disjunction can generally take scope over operators which c-command them in surface syntax. In the case of a sentence such as (22), repeated in (23) below, a wide-scope construal of disjunction results in the reading given in (24).

(23) Exactly one letter is connected with its blue circle or with its red circle.

(24) Wide scope reading: Exactly one letter is connected with its blue circle or exactly one letter is connected with its red circle.

Now, it is clear that this reading is quite marginal. Interestingly, though, some instances of the *false* condition (in which the literal, the global and the local reading are all false) happen to be cases where (24) is true. If the wide-scope reading is available to some extent, then we might expect such cases to be rated higher than all the other cases that instantiate the *false* condition. This is in fact the case: the sub-cases of the **FALSE** condition where the wide scope reading is true are judged higher than the other sub-cases of the **FALSE** condition (20% vs. 6%: $W = 128$, $p < .001$).[41]

[41]Notice that the situation is flat in the case of 'some'. In the case of 'or', we can distinguish between the sub-cases of the **FALSE** condition that make the wide-scope reading true, and the other instances of the **FALSE** condition, which make it false. Let us call the two resulting sub-conditions **FALSE [WIDE-SCOPE TRUE]** and **FALSE[WIDE-SCOPE FALSE]**. In the case of 'some', there is no well defined 'wide-scope' reading (because the quantifier phrase headed by 'some' contains a pronoun bound by 'exactly one', which as such cannot escape its scope). We can nevertheless define the **FALSE[WIDE-SCOPE TRUE]** condition in the case of 'some' as including the cases which most closely correspond to the **FALSE[WIDE-SCOPE TRUE]** condition in the case of 'or'. Once this is done, it turns out that the interaction between Scalar Item ('some' *vs.* 'or') and "**FALSE[WIDE-SCOPE TRUE]** *vs.* **FALSE[WIDE-SCOPE FALSE]**" is significant: $F(1, 15) = 11$, $p < .005$. See Appendix 3.2 for more details.

5.5.6 What do graded judgments reflect?

In section 4.4.7, we discussed the possibility that graded judgments reflect how close a given picture is perceived to be to some prototypical situation determined by the sentence it is paired with. The results of Exp. 2 do not seem to us to be be amenable to an explanation based only on typicality considerations. Indeed, there is no intuitively natural metric over the set of pictures which could explain the precise way in which our conditions are ranked by the subjects (i.e. the fact that the **ALL** condition is rated the highest, followed by the **LOCAL** and the **LITERAL** conditions). This is not to say that typicality plays no role in such a graded truth-value judgment task. As mentioned in section 4.4.7, it is perfectly plausible that the mean rating of a given condition reflects how 'close' the relevant picture is perceived to be to the typical instances of *each* available reading. What seems clear is that our results can be interpreted as reflecting typicality judgments *only if* typicality is construed as relative to *several* distinct readings.

Furthermore, the results reported in the previous subsection provide independent evidence for our general interpretation of the subjects' use of graded judgments, according to which the mean rating of a given condition is (in part) a function of the set of readings that are true in this condition (more specifically, a condition X is rated higher than a condition Y if the readings that X makes true properly include the readings that Y makes true).

5.6 Experiment 2: summary

The main result of this experiment is that scalar items in non-monotonic environments give rise to robust local readings, even more robust than the literal reading.[42] Importantly, no globalist theory of scalar implicatures can predict the local reading to be possible in such cases, where the local reading is logically independent of the literal meaning. This result thus seems to vindicate the localist approach to scalar implicatures.

6 Conclusions

We showed, first, that sentences in which a scalar item is embedded under a universal quantifier can be interpreted according to what we called the 'local' reading, contrary to Geurts and Pouscoulous' 2009 conclusions. We also clarified the theoretical status of this result and pointed out that it does not establish the existence of embedded scalar implicatures (in these cases the local reading is also predicted by some *globalist* accounts). Hence, in our second experiment, we focussed on a type of sentence where the local

[42] Notice that Geurts and Pouscoulous (2009) also found some preliminary evidence for local readings in non-monotonic environments, but disregarded it.

reading cannot be derived by globalist means, namely sentences in which a scalar item occurs in a non-monotonic environment. We were able to detect experimentally local readings for such cases.

Several questions arise. In particular, we may ask which specific aspects of our experimental design which allowed us to detect readings that Geurts and Pouscoulous (2009) did not? Note that our experimental design differed from that of Geurts and Pouscoulous in several respects. As already discussed, our pictures and task were difference (graded judgements *vs.* binary or ternary judgments). However, there were other differences as well. For instance, Geurts and Pouscoulous never presented their subjects with pictures that make the local reading true. In both our experiments, there were conditions in which the local reading is true, as well as conditions where it is false (while some other reading is true). This as such may have increased the salience of the local reading, and may therefore have contributed to the fact that the relevant sentence did not get the maximal score in cases where the local reading is false.[43] It would be interesting to find out what is the respective impact of each of these modifications on the detection of local implicatures.

Another interesting methodological question is the following: does our paradigm provide us with a general technique for detecting ambiguities, including ambiguities which speakers are not aware of? On the one hand, Geurts and Pouscoulous asked their subjects whether they perceived the relevant sentences as ambiguous and obtained a clear negative answer. On the other hand, we claim that the subjects' behavior in a graded judgement task reflected the fact that the relevant sentence is ambiguous. What we do not know is whether we managed to detect ambiguities that speakers are not aware of: contrary to Geurts and Pouscoulous, we did not ask our subjects whether they perceived the relevant sentences as ambiguous, and therefore we cannot reach a firm conclusion on this point. It would be interesting to extend this technique to other kinds of ambiguity, such as, for instance, scope ambiguities, and try to find cases where an ambiguity is not consciously perceived and yet is reflected in the subjects' answers in a graded truth-value judgement task.

Last but not least, do these results provide decisive evidence for a grammatical approach to scalar implicatures? The existence of embedded scalar implicatures is a clear prediction of grammatical approaches to scalar implicatures, and is unexpected in the neo-Gricean framework. As such, it provides an argument for grammatical theories of scalar implicatures. Nevertheless, it is not in principle excluded that a genuinely pragmatic treatment of embedded scalar implicatures could be given. As pointed out by Geurts (2009), such a treatment would have to depart significantly from the traditional Gricean approach to SIs, and would have to resort to enrichment mechanisms which, though 'local', could nevertheless be considered 'pragmatic'. Geurts (2009) suggests a

[43] Thanks to a reviewer for relevant discussions.

mechanism of 'reconstrual', similar to what Recanati (2004) calls 'free enrichment'. Testing such hypotheses would require that they be made formally explicit. Within the grammatical approach, one remaining task is to formulate explicit constraints regarding the syntactic distribution of embedded SIs (see Fox and Spector 2008 and Singh (2008) for proposals and Chemla 2009c for new empirical challenges).

Authors' addresses

Emmanuel Chemla
LSCP
Ecole Normale Supérieure
29 rue d'Ulm – Pavillon Jardin
75005 Paris, France
e-mail: chemla@ens.fr

Benjamin Spector
Institut Jean Nicod
Ecole Normale Supérieure
29 rue d'Ulm – Pavillon Jardin
75005 Paris, France
e-mail: benjamin.spector@ens.fr

Acknowledgements

We wish to thank Danny Fox, Bart Geurts and Philippe Schlenker for very useful comments. Many thanks also to Thomas Andrillon, Vincent Berthet, Isabelle Brunet, Paul Égré, Anne-Caroline Fievet, Greg Kobele, Inga Vendelin, to audiences at the University of Maryland in October 2009, at UCLA in February 2010 and of classes taught at Ealing 2009 and at the University of Vienna in June 2009. We are very grateful to the editor and anonymous reviewers for Journal of Semantics, whose very detailed comments helped us improve this paper in important ways. This work was supported by a 'Euryi' grant from the European Science Foundation ("Presupposition: A Formal Pragmatic Approach") and by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement #229 441 – CCC.

Appendix 1: Instructions (translated from French)

Thank you for your participation to this experiment. You are going to see sentences and situations in which letters are surrounded with a given number of circles. There may be connections (lines) between a letter and its circles. Here is an example:

⟨Fig. 3a for experiment 1 – 6-grid replaced with 3-grid for experiment 2⟩

Your task is to tell whether the sentence is true or false in this situation. For instance, in the example above, the sentence is true.

Important: this is not a math test! In fact, in many cases that you will see, the sentence will not be clearly true or clearly false, but it will describe more or less appropriately the situation, will be more or less natural in this situation. Hence, there is no good answer, and we are mainly interested in **your intuition**. For this reason, you can give more fine-grained judgments than a simple "yes" or "no": you will give your answers by setting the length of a red line along a line from "No" to "Yes". The more the sentence seems true/appropriate, the more you will extend the line to the right with the mouse, close to "Yes". This will certainly be the case for examples like the one above and your answer should thus look like what was represented in the frame above. On the contrary, if the sentence seems rather inappropriate to you, you will move the extremity of the line towards the left.

Let us take another example:

⟨Fig. 3b for experiment 1 – 6-grid replaced with 3-grid for experiment 2⟩

People presented with this sentence in this context have different judgments and many hesitate between "Yes" and "No". This is certainly because this sentence can be interpreted in different ways, ways which are more or less vague: "Every letter and every circle are connected" or "there are connections between letters and circles" etc. You might hesitate, but follow your feeling and the more the sentence seems spontaneously inappropriate in this situation, the more you would answer close to "No", as represented above.

In short, use the flexibility of the red bar to represent your intuition about the correspondence between the sentence and the situation. Do not try to motivate your answer or to understand where your intuition may come from: answer as you judge appropriate!

This experiment is not long, but it is a bit repetitive and you have to stay focussed. Even if you need to be attentive, do not spend too much time on each question, follow your intuition: read the sentence naturally **as if it was uttered** and, considering the picture, report your intuition using the red judgment bar. You will get used quickly and intuitively to this bar.

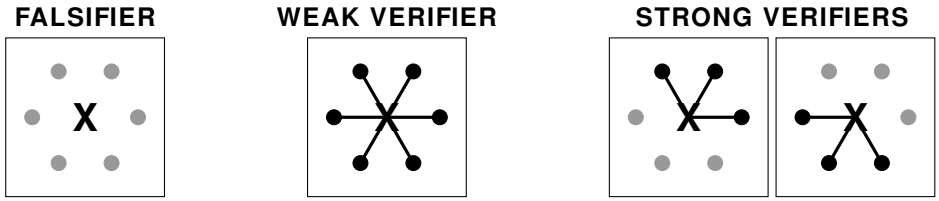
Sometimes the questions look like each other; this does not matter, always answer following your intuition for each example, independently of your previous answers.

Click 'Start' when you are ready.

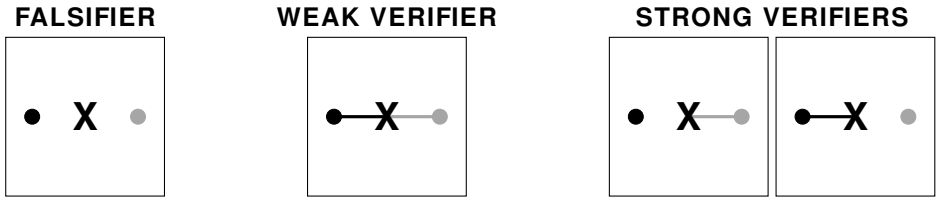
Appendix 2: Our experimental pictures

Each image used in the present experiments was a grid with 3 or 6 cells. Each cell consisted of a letter (ranging from A to C or from A to F) surrounded with circles. We can distinguish 4 types of cells: (i) **FALSIFIERS**: cells which make the predicate false, (ii) **WEAK**

VERIFIERS: cells which make the predicate true under its bare literal meaning but not under its strong interpretation, (iii) **STRONG VERIFIERS**: cells which make the predicate true under its strong interpretation. See Fig. 14 for actual examples.



(a) Possible configurations of each cell for items with the scalar item 'some'. The relevant predicate is 'x is connected with some of its circles', see full sentence in (8).



(b) Possible configurations of each cell for items with the scalar item 'or'. The relevant predicate is 'x is connected with its red circle or with its blue circle, see full sentence in (9)'.

Figure 14: Possible configurations for a letter and its circles. The predicate can be either false (**FALSIFIER**), true under its weak reading only (**WEAK VERIFIER**), or true under its strong reading (**STRONG VERIFIER**).

Notice that strong verifiers may be instantiated with different visual configurations depending on which subset of circles ends up being connected to the letter. Hence, each condition in which there were several strong verifiers was duplicated: in one version, the strong verifiers were all the same (this situation is referred to with an $=$ sign), in another version, there were instantiated with two different configurations (referred to with $\neq$). The potential importance of this specific variation is discussed in section 4.4.5.

Appendix 2.1: Target conditions, experiment 1

The pictures instantiating the target conditions for experiment 1 were grids with 6 cells constructed from the building blocks described above and in Fig. 14. Table 1 summarizes the list of items according to the number of falsifiers, weak verifiers and strong verifiers the grid contained and indicates for each reading whether it is true ($\checkmark$) or false ($*$).

	FALSE-0	FALSE-2	FALSE-4	LITERAL	WEAK-2		WEAK-4		STRONG	
# falsifiers:	6	4	2	0	0	0	0	0	0	0
# weak verifiers:	0	0	0	6	4	4	2	2	0	0
# strong verifiers:	0	2	4	0	2	2	4	4	6	6
Strong verifiers are:		$\neq$	$\neq$		$=$	$\neq$	$=$	$\neq$	$=$	$\neq$
Literal reading:	$*$	$*$	$*$	✓	✓	✓	✓	✓	✓	✓
Global reading:	$*$	$*$	$*$	$*$	✓	✓	✓	✓	✓	✓
Local reading:	$*$	$*$	$*$	$*$	$*$	$*$	$*$	$*$	✓	✓

Table 1: List of conditions for the target sentences (8) and (9) in Exp. 1. See Fig. 14 for a definition of 'falsifiers' and 'weak/strong verifiers'. Each kind of picture was instantiated twice for each scalar item in each block.

Appendix 2.2: Control DE items, experiments 1 and 2

The list of pictures used to instantiate the control conditions described in section 4.2.2 and 5.3.2 is shown in Table 2. Pictures are described according to the number of falsifiers, weak verifiers and strong verifiers.

	FALSE		?LOCAL	BOTH
# falsifiers:	0	0	0	6
# weak verifiers:	0	0	6	0
# strong verifiers:	6	6	0	0
Strong verifiers are:	$\neq$	$=$		
"Local" reading:	$*$	$*$	✓	✓
Literal reading:	$*$	$*$	$*$	✓

Table 2: List of conditions for the control sentences (12) and (13) in Exp. 1 and 2. See Fig. 14 for a definition of 'falsifiers' and 'weak/strong verifiers'. Each kind of picture was instantiated twice for each scalar item except for the one corresponding to the **BOTH** condition which was instantiated 4 times for each scalar item.

Appendix 2.3: Target conditions, experiment 2

As before, these sentences were presented with grids of items which falsified or verified the weak or the strong interpretation of the embedded predicate (see Fig. 14). Contrary to

experiment 1, each grid now contained only 3 cells, the detailed description of which can be found in Table 3.

	FALSE									LOCAL		LITERAL	ALL
# falsifiers:	3	0	1	0	0	0	0	1	1	0	1	2	2
# weak verifiers:	0	2	2	0	0	1	1	0	0	2	1	1	0
# strong verifiers:	0	0	0	3	3	2	2	2	2	1	1	0	1
Strong verifiers are:				=	≠	=	≠	=	≠				
Local reading:	*	*	*	*	*	*	*	*	*	✓	✓	*	✓
Literal reading:	*	*	*	*	*	*	*	*	*	*	*	✓	✓
Global reading:	*	*	*	*	*	*	*	*	*	*	*	*	✓
wide scope:	*	*	*	*	✓	*	✓	*	✓	*	✓	*	✓

Table 3: List of conditions for the target sentences in experiment 2. See Fig. 14 for a definition of 'falsifiers' and 'weak/strong verifiers'. Each kind of picture was instantiated twice for each scalar item in each block. We indicated in the last row whether the sentence is true when disjunction takes wide scope over the quantifier (cf. section 5.5.5—we ignore the potential implicature of the resulting construction).

Appendix 3: Fine-grained results for experiments 1 and 2

Appendix 3.1: Fine-grained results for experiment 1

In Fig. 15, we report more fine-grained results for sub-conditions of experiment 1 as described in 2.1. (The **FALSE-0/2/4** and **LITERAL** conditions were already reported in Fig. 10.)

Appendix 3.2: Fine-grained results for experiment 2

In Fig. 16, we report more fine-grained results for sub-conditions of experiment 2 (see 2.3). We annotated some sub-conditions of the **FALSE** condition with the mention 'wide-scope', to indicate that those sub-conditions were the only ones, within the **FALSE** condition, which satisfy the marginally available reading in which disjunction takes scope over 'exactly one' (in the case of the conditions involving disjunction, cf. section 5.5.5).

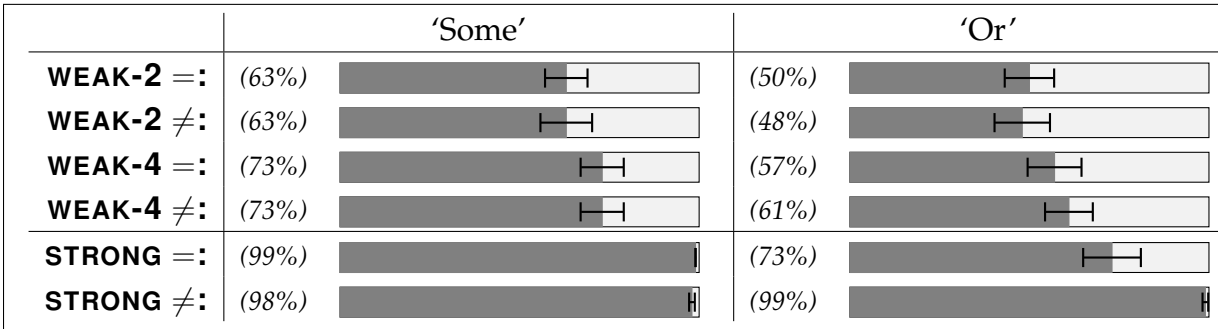


Figure 15: Fine-grained results for sub-conditions of exp. 1 as described in 2.1. See Fig. 10 for the **FALSE-0/2/4** and **LITERAL** conditions.

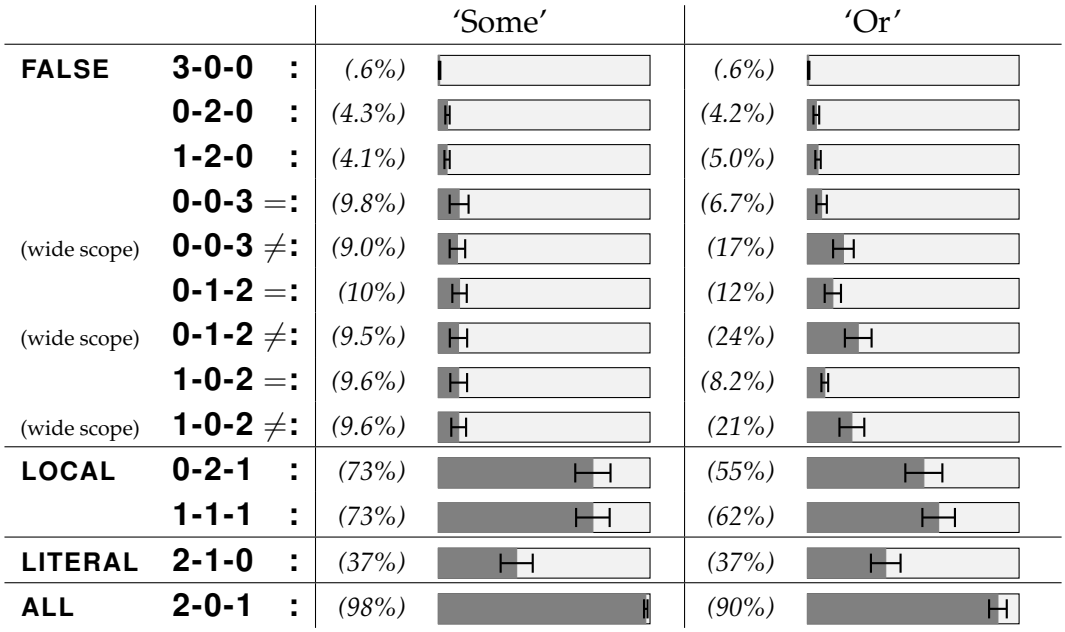


Figure 16: Fine-grained results for sub-conditions of exp. 2 as described in 2.3: **X-Y-Z** describes the numbers of falsifiers (x), weak verifiers (y) and strong verifiers (z).

References

Abusch, Dorit. (1993). The scope of indefinites. *Natural Language Semantics* 2:83–135.

Armstrong, Sharon Lee, Lila R. Gleitman, and Henry Gleitman. (1983). What some concepts might not be. *Cognition* 13:263–308.

Bard, Ellen Gurman, Dan Robertson, and Antonella Sorace. (1996). Magnitude estimation of linguistic acceptability. *Language* 72:32–68.

Chemla, Emmanuel. (2008). Présuppositions et implicatures scalaires: études formelles et expérimentales. Doctoral Dissertation, ENS.

Chemla, Emmanuel. (2009a). Presuppositions of quantified sentences: experimental data. *Natural Language Semantics* 17:299–340.

Chemla, Emmanuel. (2009b). Similarity: towards a unified account of scalar implicatures, free choice permission and presupposition projection. Under revision for *Semantics and Pragmatics*.

Chemla, Emmanuel. (2009c). Universal implicatures and free choice effects: Experimental data. *Semantics and Pragmatics* 2:1–33.

Chemla, Emmanuel, and Philippe Schlenker. (2009). Incremental vs. symmetric accounts of presupposition projection: An experimental approach. Ms. IJN, LSCP & NYU.

Chemla, Emmanuel, and Benjamin Spector. (2010). Experimental evidence for embedded implicatures. In Maria Aloni and Katrin Schulz (eds.), *Proceedings of the amsterdam colloquium 2009*. Springer.

Chierchia, Gennaro. (2004). Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. In A. Belletti (ed.), *Structures and beyond*. Oxford University Press.

Chierchia, Gennaro. (2006). Broaden Your Views: Implicatures of Domain Widening and the 'Logicality' of Language. *Linguistic Inquiry* 37:535–590.

Chierchia, Gennaro, Danny Fox, and Benjamin Spector. (in press). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics.

Cohen, Jonathan L. (1971). Some Remarks on Grice's Views about the Logical Particles of Natural Language. In Bar Hillel, Yehoshua (ed.), *Pragmatics of natural languages*, 50–68. Reidel Dordrecht.

Cowart, Wayne. (1997). *Experimental syntax: applying objective methods to sentence judgments*. Thousand Oaks, CA.

Crain, Stephen, and Rosalind Thornton. (2000). *Investigations in Universal Grammar: A guide to experiments on the acquisition of syntax and semantics*. The MIT Press.

Dalrymple, Mary, Makoto Kanazawa, Yookyung Kim, Sam Mchombo, and Stanley Peters. (1998). Reciprocal expressions and the concept of reciprocity. *Linguistics and Philosophy* 21:159–210.

Davidson, Donald. (2001). *Inquiries into truth and interpretation*. Oxford University Press, USA.

Fox, Danny. (2007). Free Choice and the theory of Scalar Implicatures. In Uli Sauerland and P. Stateva (eds.), *Presupposition and Implicature in Compositional Semantics*, 537–586. New York, Palgrave Macmillan.

Fox, Danny, and Benjamin Spector. (2008). Economy and embedded exhaustification. Handout.

Fritzley, V. Heather, and Kang Lee. (2003). Do young children always say yes to yes-

no questions? A metadevelopmental study of the affirmation bias. *Child Development* 74:1297–1313.

Geurts, Bart. (2009). Scalar implicature and local pragmatics. *Mind & Language* 24:51–79.

Geurts, Bart, and Nausicaa Pouscoulous. (2008). No scalar inferences under embedding. In Paul Egré and Giorgio Magri (eds.), *Presuppositions and implicatures.*. MIT Working Papers in Linguistics.

Geurts, Bart, and Nausicaa Pouscoulous. (2009). Embedded implicatures?!? *Semantics and Pragmatics* 2:1–34.

Grice, H. Paul. (1967). Logic and conversation. *the William James Lectures, delivered at Harvard University. Republished in Grice (1989)* .

Grice, H. Paul. (1989). *Studies in the way of words.*. Harvard University Press, Cambridge, Massachusetts.

Groenendijk, Jeroen A.G., and Martin J.B. Stokhof. (1984). Studies in the semantics of questions and the pragmatics of answers. Doctoral Dissertation, University of Amsterdam.

Gualmini, Andrea, Stephen Crain, Luisa Meroni, Gennaro Chierchia, and Maria Teresa Guasti. (2001). At the semantics/pragmatics interface in child language. In *Proceedings of Semantic and Linguistic Theory 11 (SALT 11)*.

Hale, John. (2001). A probabilistic earley parser as a psycholinguistic model. In *Naacl '01: Second meeting of the north american chapter of the association for computational linguistics on language technologies 2001*, 1–8. Morristown, NJ, USA: Association for Computational Linguistics.

Horn, Laurence R. (1985). Metalinguistic negation and pragmatic ambiguity. *Language* 121–174.

Horn, Laurence R. (2006). The border wars: A neo-Gricean perspective. *Where semantics meets pragmatics* 21–48.

Klinedinst, Nathan. (2006). Plurality and possibility. Doctoral Dissertation, UCLA.

Kratzer, Angelika, and Junko Shimoyama. (2002). Indeterminate pronouns: The view from Japanese. *The Proceedings of the Third Tokyo Conference on Psycholinguistics* 1–25.

Landman, Fred. (1998). Plurals and Maximalization. In S. Rothstein (ed.), *Events and grammar*, 237–271. Kluwer, Dordrecht.

Levinson, Stephen C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. MIT Press Cambridge, MA, USA.

Magri, Giorgio. (2009). A theory of individual level predicates based on blind mandatory implicatures. *Natural Language Semantics* 17:245–297.

Meyer, Marie-Christine, and Uli Sauerland. (2009). A pragmatic constraint on ambiguity detection. *Natural Language & Linguistic Theory* 27:139–150.

Moriguchi, Yusuke, Mako Okanda, and Shoji Itakura. (2008). Young children's yes bias: How does it relate to verbal ability, inhibitory control, and theory of mind? *First Lan-*

guage 28:431.
Noveck, Ira A. (2001). When children are more logical than adults: experimental investigations of scalar implicature. *Cognition* 78:165–188.
Quine, Willard Van Orman. (1964). *Word and object*. MIT press.
Recanati, François. (2003). Embedded Implicatures. *Philosophical Perspectives* 17:299–332.
Recanati, François. (2004). *Literal meaning*. Cambridge University Press.
Reinhart, Tanya. (1997). Quantifier scope: How labor is divided between QR and choice functions. *Linguistics and Philosophy* 20:335–397.
van Rooij, Robert, and Katrin Schulz. (2004). Exhaustive Interpretation of Complex Sentences. *Journal of Logic, Language and Information* 13:491–519.
Sauerland, Uli. (2004). Scalar Implicatures in Complex Sentences. *Linguistics and Philosophy* 27:367–391.
Sauerland, Uli. (2005). The epistemic step. *Experimental Pragmatics, University of Cambridge, April* .
Sauerland, Uli. (2010). Embedded implicatures and experimental constraints: A reply to Geurts & Pouscoulous and Chemla. *Semantics and Pragmatics* 3:1–13.
Schulz, Katrin. (2003). You may read it now or later. A case study on the paradox of free choice permission. Master's thesis, University of Amsterdam.
Schütze, Carson T. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. University of Chicago Press.
Singh, Raj. (2008). Modularity and locality in interpretation. Doctoral Dissertation, MIT.
Spector, Benjamin. (2003). Scalar implicatures: Exhaustivity and Gricean reasoning. In Balder ten Cate (ed.), *Proceedings of the eigth esslli student session*. Vienna, Austria. Revised version in Spector (2007b).
Spector, Benjamin. (2006). Aspects de la pragmatique des opérateurs logiques. Doctoral Dissertation, Université Paris 7.
Spector, Benjamin. (2007a). Aspects of the Pragmatics of Plural Morphology: On Higher-order Implicatures. In Uli Sauerland and Penka Stateva (eds.), *Presuppositions and Implicatures in Compositional Semantics*, 243–281. Palgrave Macmillan New York.
Spector, Benjamin. (2007b). Scalar implicatures: Exhaustivity and Gricean reasoning. In Maria Aloni, Paul Dekker, and Alastair Butler (eds.), *Questions in dynamic semantics*, Vol. 17 of *Current Research in the Semantics/Pragmatics Interface*, 225–249. Elsevier.
Stevens, Stanley Smith. (1956). The direct estimation of sensory magnitudes-loudness. *American Journal of Psychology* 69:1–25.