The why and how of experimental pragmatics: The case of 'scalar inferences'

Ira Noveck & Dan Sperber

In press, *Advances in Pragmatics (ed. Noel Roberts), Palgrave*

Although a few pioneers in psycholinguistics had, for more than twenty years, approached various pragmatic issues experimentally, it is only in the past few years that investigators have begun employing the experimental method in testing pragmatic hypotheses (see Noveck & Sperber 2004). We see this emergence of a proper experimental pragmatics as an important advance with a great potential for further development. In this chapter we want to illustrate what can be done with experimental approaches to pragmatic issues by presenting one case, that of so-called 'scalar inferences', where the experimental method has helped sharpen a theoretical debate and has provided uniquely relevant evidence. We will focus on work done by the first author and his collaborators or work closely related to theirs, but other authors have also made important contributions to the topic (e.g. Papafragou and Musolino, 2003; Guasti, Chierchia, Crain, Foppolo, Gualmini, & Meroni, 2005; De Neys & Schaeken, in press).

Methodological background: the limits of pragmatic intuitions as evidence

Theoretical work in pragmatics relies heavily – often exclusively – on pragmatic intuitions. These are rarely complemented with observational data of a kind more common in sociologically oriented pragmatics. The use of statistical data from corpuses and from experiments is even less common. This situation results partly from the fact that most theoretical pragmatists have been trained in departments of linguistics where, quite often, linguistic intuitions are the only kind of data considered. Optimally, of course, one would want pragmatists to use whatever kind of data that may significantly confirm or disconfirm hypotheses. Moreover, a sensible methodological pluralism is not the only reason to diversify the types of evidence used in pragmatics. There are also principled limits to the use of pragmatic intuitions.

It makes sense (even if it is not uncontroversial) to judge a semantic description by its ability to account for semantic intuitions. Of course, the use of semantic intuitions and of linguistic intuitions generally, raises methodological problems and calls for methodological caution. For instance, a linguist's intuitions may be biased by prior theoretical commitments. Also, one may mistake what are in fact pragmatic intuitions for semantic ones (as, Grice argued, ordinary language philosophers systematically did). Still, there are good reasons why semantic intuitions are so central to semantics. Semantic intuitions are not just *about* semantic facts; they are semantic facts themselves. For instance, the intuition that sentence (1) entails

(2) is not *about* some semantic property that this sentence would have anyhow, regardless of its accessibility to speakers' intuitions.

(1) John knows that it is raining
(2) it is raining

Rather, for (1) to have the meaning it has *is*, among other aspects, to be intuitively understood as entailing (2). A semantic analysis of linguistic expressions that accounts for all the speaker-hearer's semantic intuitions about these expressions may not be the best possible analysis, but it is descriptively adequate (in Chomsky's sense – an explanatory adequate description of the semantics of a given language, on the other hand, involves hypotheses about the capacities that make the acquisition of this semantics possible, and here observational and experimental evidence should be of relevance).

The use of pragmatic intuitions raises the same methodological problems as does the use of semantic intuitions and then some. It is a mistake to believe that pragmatic intuitions of the kind used in pragmatics are data of the same kind as semantic intuitions used in semantics. Genuine pragmatic intuitions are those that addressees have about the intended meaning of an utterance addressed to them. Quite generally, pragmatic intuitions invoked in theoretical pragmatics are not about actual utterances addressed to the reader of a pragmatic article, but about hypothetical cases involving imaginary or generic interlocutors. Pragmatic intuitions on hypothetical utterances have proved useful in a variety of ways, but it is important to keep in mind that these are not about how an utterance is interpreted, but about how an utterance *would be* interpreted if it were produced in a specific situation by a speaker addressing a listener, with referring expressions having actual referents, and so on. These intuitions are educated guesses – and, no doubt, generally good ones – about hypothetical pragmatic facts, but are not themselves pragmatic facts and they may well in error. That is, we may be wrong about how, in fact, we would interpret a given utterance in a given context.

Besides helping compensate for the inherent limits of pragmatic intuitions, an experimental approach can provide crucial evidence when deciding between alternative theories that may agree on the content of the interpretations of utterances, but that have different implications regarding the cognitive mechanisms through which these interpretations are arrived at. Of course, for their contribution to be of value, experimentalists must conform to fairly strict methodological criteria and measure just what they are intent on measuring—

typically the effect of one 'independent' variable on another 'dependent' variable without other uncontrolled variables affecting the results. We will show how this plays out in the study of 'scalar inferences'.

Theoretical background: Scalar implicatures as Generalised Conversational Implicatures (GCIs)

The experiments we will present are relevant to the study of so-called 'scalar implicatures.' Here we just remind readers of the main features of the Gricean and neo-Gricean account of scalar implicatures, and focus on the claim that scalar implicatures are Generalized Conversational Implicatures, or GCIs. Scalar implicatures are illustrated by cases such as (3a) which is said to implicate (3c), or (4a) said to implicate (4c):

(3) (a) It is possible that Hillary will win
(b) It is certain that Hillary will win
(c) It is not certain that Hillary will win
(4) (a) Some of the guests have arrived
(b) All of the guests have arrived
(c) Not all of the guests have arrived

Proposition (3b) is more informative than (3a), which it entails. If the more informative proposition would make a greater contribution to the common purpose of the conversation, then, a speaker obeying Grice's first Maxim of Quantity ("Make your contribution as informative as is required") would be expected to express it unless she could not do so without violating the Supermaxim of Quality ("Try to make your contribution one that is true"). Hence, on a Gricean account, a speaker stating (3a) typically implicates (3c) (i.e., the negation of (3b)). For the same reasons, a speaker stating (4a) typically implicates (4c) (i.e., the negation of (4b)).

Such implicatures are described as 'scalar' because, according to an account developed by neo-Griceans and in particular Lawrence Horn (1972), the derivation of these implicatures draws on pre-existing linguistic scales consisting in a set of alternate terms or expressions ranked by order of informativeness; <*possible, certain*> and <*some, all*> are examples of such scales. When a less informative term is used in an utterance in a way that does not satisfy the first maxim of quantity, the speaker can be taken to implicate that the proposition that would

have been expressed by the use of a stronger term in the scale is false. This account of implicatures such as those carried by (3a) or (4a) extend to a wide variety of cases and has some intuitive appeal. It should not be seen however as obviously correct or without alternatives. In particular, its implications for processing are less attractive. According to such an account, the inference from the utterance to its scalar implicature goes through a consideration not just of what the speaker said and the context but also of what the speaker might have said but did not. It is this type of onerous inference that makes the Gricean account of implicature derivation seem implausible from a cognitive and developmental point of view.

Levinson draws on another idea of Grice, that of Generalized Conversational Implicatures, to offer an account that might provide a solution to the problem posed by the derivational complexity of scalar implicatures. Grice noted that some implicatures are generally valid (from a pragmatic rather than logical point of view, of course) and therefore could be inferred without consideration of the context, except in cases where the context happens to make them invalid. Grice contrasted these Generalized Conversational Implicatures with Particular Conversational Implicatures, which are valid only in specific contexts. In his book *Presumptive Meanings: The Theory of Generalized Conversational Implicatures* (Levinson 2000), Levinson elaborates Grice's original and somewhat vague notion. For Levinson, GCIs are *default inferences*, that, is, inferences that are automatically generated and that can be cancelled if there are contextual reasons to do so. Levinson treats scalar implicatures as paradigmatic cases of GCIs (whereas Grice's own examples of GCIs don't include scalar implicatures). This has the advantage of making the inference of these implicatures a relatively light one-step process, which needs to access neither contextual premises nor the full Gricean rationale for their derivation.

Levinson's own rationale for GCIs so conceived has to do with the optimization of processing. The existence of GCIs speeds up the process of communication that is slowed down, Levinson argues, by the need for phonetic articulation: some unencoded aspects of the speaker's meaning can be inferred from metalinguistic properties of the utterance such as the choice of a given word from among a set of closely related alternatives. For instance the speaker's choice of "some" rather than the stronger "all" in (4a) ("Some of the guests have arrived") justifies inferring that (4c) is part of her meaning. These are non-demonstrative inferences, or course. There are cases where these inferences are invalid. For instance, if it were contextually established that the speaker of (4a) has only partial information about the

arrival of the guests, then (4c) would not be part of her meaning. Still, given that GCIs are valid in most contexts (or so it is assumed), the overall speeding up of communication made possible by the automaticity of GCIs is not compromised by the rare cases where contextual considerations force the hearer to countermand them.

The theory of scalar implicatures as default GCIs combines four claims:

(a) These inferences are made by default, irrespective of the context, and cancelled when the context demands

(b) The fact that these inferences are made by default adds to the speed and efficiency of communication

(c) These inferences contribute implicatures to the interpretation of the utterance, as opposed to contributing enrichments of its explicit content ('what is said' in Grice's terms or 'explicatures' in relevance theory's terms)

(d) These inferences are scalar: they exploit pre-existing scales such as <*some*, *all*>, <*or*, *and*>, <*possible*, *necessary*>

We doubt all four claims. The bulk of this chapter will be devoted to explaining how experimental evidence has cast strong doubts on claim (a). First, however, we briefly present an argument that also casts doubt on (b), and we outline the relevance-theoretic approach, which is in contradiction with all four claims.

This idea that default implicatures or GCIs would permit more efficient and speedier communication may seem sensible and capable of lending support to the whole theory. It raises however the following empirical issue. If the frequency of GCI cancellations were too high, their cost would offset the benefit of deriving GCIs by default. Suppose for instance that a given type of GCI had to be cancelled a third of the time. The cost of the use of such a GCI would be that of deriving it by default in all cases plus the cost of cancelling it in one third of the cases. This would have to be compared with the cost of deriving the implicature as a 'particularized conversational implicature,' that is, in a contextually sensitive and therefore more costly way, in two thirds of the cases, but without any cost of default derivation followed by cancellation in the other third of the cases. It is not clear that, with such frequencies, the rationale given for GCIs in term of economy would make much sense.

To show that this kind of calculus is not unrealistic, consider the example of "P or Q" and its alleged GCI *not (P and Q)*. We are not aware of any statistical data regarding the frequency of exclusive uses of "or" and we share the common intuition that quite often, when people utter a sentence of the form "P or Q" they can be taken to consider that *P and Q* is

excluded. This exclusion however need not be part of their meaning. In most cases this exclusion follows from real world knowledge and not from the interpretation of "or", as illustrated in (5)-(7):

(5) He is a bachelor or he is divorced
(6) Jane is in Paris or in Madrid
(7) Bill will arrive Monday or Tuesday

If "P or Q" implicates by default that *not (P and Q)*, then, in all cases such as (5)-(7) where the two disjuncts cannot both be true for commonsense reasons, people automatically compute a GCI that causes the speaker's meaning to redundantly implicate what is already part of the common ground, and surely, this is a cost without associated benefit. Moreover, if one carefully excludes cases where mutual exclusivity of the disjuncts is self-evident and need not be communicated, and looks at cases such as (8)-(10) where neither the inclusive nor the exclusive interpretation is a priori ruled out, it is not at all obvious that the exclusive interpretation of "or" is dominant:

(8) She wears sunglasses or a cap
(9) Our employees speak French or Spanish
(10) Bill will sing or play the piano

We have no hard statistical data to present, but it seems less than obvious that a disposition to understand by default utterances of the form "P or Q" as implicating *not (P and Q)* would render communication speedier or more efficient. More generally, the effect that GCIs would have on the efficiency of communication should be investigated rather than assumed.

Relevance theory's approach

We will assume that the basic tenets of relevance theory are familiar (see Wilson & Sperber 2004 for a recent restatement), and focus on how it applies to what neo-Griceans describe as 'scalar implicatures.' Two basic ideas play a crucial role here:

(a) Linguistic expressions serve not to *encode* the speaker's meaning but to *indicate* it. The speaker's meaning is inferred from the linguistic meaning of the words and expressions used taken together with the context.

(b) Inferring the speaker's explicit and implicit meaning (her explicatures and implicatures) is not done sequentially but in parallel. The final overall interpretation of an utterance results from a mutual adjustment of implicatures and explicatures guided by expectations of relevance.

Here is a simple illustration of these two points:

(11) *Henry*: Do you want to go on working, or shall we go to the cinema?
Jane: I am tired. Let's go to the cinema.

Jane's describing herself as "tired" achieves relevance as an explanation of her acceptance of Henry's suggestion. For this it must be understood that she is not just tired, but too tired to go on working, and at the same time not too tired to go to the cinema. Her used of "tired" serves to indicate an ad hoc concept TIRED* with an extension narrower than that of the linguistically encoded concept TIRED. Whereas TIRED extends from a minimal level of tiredness to complete exhaustion, TIRED* extends just over those levels of tiredness that explain why Jane would rather go to the cinema than work. Henry correctly understands Jane explicature to be (12) and her implicature to be (13), yielding an optimally relevant interpretation:

(12) I am TIRED*
(13) The reason why I would rather go to the cinema than work is that I am TIRED*

Note that explicature (12), and in particular the interpretation of "tired" as indicating TIRED* is calibrated so as to justify implicature (13). The explicature therefore could only be inferred once the implicature had been tentatively assumed to be part of Jane's meaning. The overall interpretation results from a process of mutual adjustment between explicature and implicature.

Consider now an expression typically supposed to give rise to 'scalar implicatures' such as "some of the Xs". From a semantic point of view, "some of the Xs" has as its extension the set of subsets of n Xs where n is at least 2 and at most the total number of the Xs. From a relevance-theoretic pragmatic point of view, the use of an expression of the form "some of the Xs", just as that of any linguistic expression, serves not to encode the speaker's meaning, but to indicate it. In particular the denotation of the concept indicated by a given use of "some of

the Xs" may be an ad hoc concept SOME OF THE Xs* with a denotation different from that of the literal SOME OF THE Xs. Rather than ranging over all subsets of Xs between 2 and the total number of the Xs, the extension of SOME OF THE Xs* may be narrowed down at either end, or it may be extended so as to include subsets of one.

Imagine (14) uttered in a discussion of the spread of scientific knowledge in America:

(14) Most Americans are creationists and some even believe that the Earth is flat

Clearly, the speaker is understood as meaning that a number of Americans much greater than two believe that the earth is flat. Two Americans with such a belief—say two inmates in a psychiatric hospital—would be enough to make her utterance literally true, but not, and by a wide margin, to make it relevant. Given that the speaker can be assumed to know that it is common knowledge that not all Americans believe that the earth is flat, there is no ground to assume that this is a part of her meaning (inferring it would not add any cognitive effect and it would involve a processing cost, hence it would detract from relevance). On the other hand, the speaker's contrastive use of "most" and "some" and her use of "even" make it part of her meaning that the Americans who believe the earth to be flat are fewer than those who believe in creationism (this, of course, entails that not all Americans believe that the earth is flat, but not every entailment of a speaker's meaning is part of that meaning). So the denotation indicated by "some" in (14) is narrower than its literal denotation at both ends: the subsets of Americans in the denotation of this occurrence of "some" are large enough to be relevant and hence much larger than sets of two Americans, and are smaller than the set of American creationists.

Let us now go back to a version of example (4). Jane and Henry have invited a few friends to a dinner party. Suppose first that it was agreed that Henry would go and get the desert from the pastry shop as soon as the guests started arriving. Henry is in the garage, he hears the bell ring and then Jane shouting (15) to him:

(15) *Jane to Henry*: Some of the guests have arrived

Henry does not know whether one, many, or all the guests have arrived, or, for that matter, whether Jane has already opened the door and seen how many of them there are, and the question need not even come to his mind. What makes Jane's utterance relevant is that it

implies that he should go now and this does not depend on the number of guests at the door. Henry's construal of "some" is compatible with any number of guests having arrived, even just one, and hence is an extended construal of "some".

Consider now a different scenario. Henry is alone in the kitchen cooking. Jane comes in and tells him (15). The consequences that Henry considers are that he should come and greet the guests and bring the finger food he has prepared as an appetizer. The value of "some" is taken to be a value for which these are the main consequences. If all the guests had come, what he should do would be not just to greet the guest and bring the finger food, but also and even more importantly, to put the fish in the oven and make the ultimate preparations for the meal itself. The fact that Jane's utterance achieves relevance without bringing to mind consequences more typical of the arrival of all the guests causes Henry to construe "some" with some vague cardinality above one and below all. Henry need not actively exclude *all*, he may just not even consider it. If however Henry is wondering whether all the guests have arrived, then he will take Jane's utterance to licence the inference that not all of them have. If moreover Henry had asked Jane whether all the guests had arrived, or if he knew that she knew that it was particularly relevant to him at this point in time, he would take that inference to be intended. He would also do so if she had put a contrastive stress on "some", causing an extra effort and suggesting an extra effect. In other words, if there is some mutually manifest, actively represented reason to wonder whether all the guests have arrived, then (15) can be taken to implicate that not all of them have.

From a relevance theory point of view, (11), (14), and (15) are just ordinary illustrations of the fact that linguistic expressions serve to indicate rather than encode the speaker's meaning and that the speaker's meanings are quite often a narrowing down or broadening of the linguistic meaning. Taking "some" to indicate not *at least two and possibly all* but *at least two and fewer than all* is a common narrowing down of the literal meaning of "some" at the level of the explicature of the utterance. It is not automatic but takes place when the consequences that render the utterance relevant as expected are characteristically carried by this narrowed down meaning.

We are not denying that a statement of the form "…some…" may, in some cases carry an implicature of the form …*not all*… (or, in other cases we will not discuss here, an implicature of the form …*some…not*…). This occurs when the "…some…" utterance achieves relevance by answering a tacit or explicit question as to whether *all* items satisfy the predicate.The fact that it does not answer it positively *implicates* a negative answer and

therefore a narrowed down construal of "some" as excluding all. Standard accounts of 'scalar implicatures' fail to distinguish the cases where the explicature merely entails …*not all.*… and the much less frequent cases where, moreover, the utterance implicates …*not all.*… .

In all the cases where the meaning of "some" in an utterance is narrowed down so as to exclude *all*, this is the result of an inferential process that looks at consequences that might make the utterance relevant as expected, and that adjusts the meaning indicated by "some" to these consequences. In particular if what may make the utterance relevant is an implication that is true of some Xs but not of all Xs, then the meaning of "some" is adjusted so as to exclude *all*. These inferential processes result from the automatic attempt by the hearer to find an interpretation of the utterance that meets his expectation of relevance and they all follow the same heuristics. There is nothing distinctive in the way 'scalar' inferences are drawn. Moreover, the class of cases described in the literature as scalar inferences is characterised by an enrichment at the level of the explicature (where, for instance, "some" is reinterpreted in a way that excludes *all*) and only in a small sub-class of these is the exclusion of the more informative concept not just entailed but also implicated.

According to relevance theory, then, so called 'scalar implicatures' are not scalar, nor necessarily implicatures. Of course, the notion of 'scalar implicature' could be redefined to fit just cases where there is an explicit or implicit question as to whether the use of a more informative expression than the one employed by the speaker (e.g. "all' instead of "some") would have been warranted, and in such cases, a denial of a more informative claim can indeed be implicated by the use of the less informative expression. However, 'scalar implicatures' in this restricted sense depend on contextual premises (linked to the fact that the stronger claim was being entertained as a relevant possibility) rather than on a context-independent scale, and are not candidates therefore for the status of GCI.

From the point of view of relevance theory then, the classical neo-Gricean theory of scalar implicatures can be seen as a mistaken generalisation of the relatively rare case where a weaker claim genuinely *implicates* the denial of a stronger claim that is contextually under consideration to the much more common case where the denotation of an expression is narrowed down so as to exclude marginal or limiting instances carrying untypical implications. For instance "possible" as in (3a) ("It is possible that Hillary will win") is often construed as excluding, on one side, mere metaphysical possibility with very low empirical probability, and, on the other side, certainty and quasi-certainty. The trimming of "possible" at both ends results in an enriched and generally more relevant meaning. Since the trimming

at the very high probability end is not different from that at the very low probability end, both should be explained in the same way, ruling out the scalar aspect of the 'scalar implicature' account, which works, if at all, only at the upper end. On the other hand, if (3a) were uttered in reply to the question: "Is it certain that Hillary will win?", then it would indeed implicate (3c) ("It is not certain that Hillary will win") because it would achieve relevance by implicitly answering in the negative a question that had been asked. From a relevance theory point of view, the two cases should be distinguished.

This is not the place to compare in detail the GCI and the relevance-theoretic approaches. We focus rather on a testable difference in prediction between them. Levinson writes: "GCI theory clearly ought to make predictions about process. But here the predictions have not yet been worked out in any detail" (Levinson 2000:370). There is however one prediction about process that follows quite directly from GCI theory since it is hardly more than a restatement of some of the tenets of the theory. According to the theory, GCIs are computed by default and are contextually cancelled when needed. Both the computation of GCIs and their cancellation are processes and therefore should take each some time and effort (even if the default character of GCI should make their computation quite easy and rapid). Everything else being equal, less effort should be expended and less time taken in the normal case where a GCI is computed and not cancelled than in the exceptional case where a GCI is first computed and then cancelled. Relevance theory predicts just the opposite pattern.

From a relevance-theoretic perspective, the speaker's meaning is always inferred, even when it consists in a literal interpretation of the linguistic expressions used. The inferences involved, however, differ in the time and effort they require. Both the sentence meaning and the context contribute to making some interpretations more easily derived than others. If only sentence meaning were involved, one should predict that the smaller the distance between it and the speaker's meaning it serves to indicate, the lesser would be the time and effort required to infer the speaker's meaning. Contextual factors, however, must be taken into account. For instance, an enriched interpretation may be primed by the context and, as a result may be easier to infer than a literal interpretation. Consider a variation of example (11):

(16) *Henry*: You look tired, let's go to the cinema"

Jane: I am tired, but not too tired to go on working

A natural interpretation of Henry's utterance involves the ad hoc concept TIRED* such that being TIRED* is a sufficient reason to stop working and not a sufficient reason to stay at home. Jane could have answered, "No, I am not tired, I'll go on working" meaning that she was not TIRED*. When Jane, rather, asserts that she *is* tired, Henry is primed to interpret "tired" as TIRED*. A relevant interpretation of Jane's whole utterance, however, imposes a broader, more literal and, in this situation, more effortful construal of the term.

Even when an enriched interpretation of an utterance is not primed, it may require less processing effort than would the literal interpretation because the contextual implications that render relevant the enriched interpretation are more easily arrived at than those that would render relevant the literal interpretation. This typically occurs with metaphorical utterances: a relevant literal interpretation is often hard or even impossible to construct.

In the absence of contextual factors that make an enriched interpretation of an utterance easier to arrive at, relevance theory predicts that a literal interpretation—which involves just the attribution to the speaker of a meaning already provided by linguistic decoding—should involve shallower processing and take less time than an enriched one—which involves a process of meaning construction. Such is the case in particular in the experiments we describe below.

The difference in prediction between GCI theory and relevance theory can be presented in table form:

Interpretation of the scalar term	GCI theory	relevance theory
literal	default enrichment + context-sensitive cancellation, ***hence slower***	no enrichment, ***hence faster***
enriched	default enrichment, ***hence faster***	context-sensitive enrichment, ***hence slower***

Table 1: contrasting predictions of GCI Theory and relevance theory regarding the speed of interpretation of scalar term (when an enriched construal is not contextually primed)

This difference in prediction between the two theories is of a type that lends itself to experimental investigation.

Methodological considerations in experimental approaches to 'scalar inferences'

In the experimental study of case of scalar inferences,[1] one has to keep four methodological considerations in mind. To begin with, one wants to be sure that a given result (whether it be the rate of responses that indicate a pragmatic enrichment or the mean reaction time associated with an enrichment) is a consequence of the experiment's intended target and not of other contextual variables. For example, one would want to be sure that the understanding of a disjunctive statement of the form *P or Q* as excluding *P and Q* is due to the pragmatic enrichment of the term "or" (from an inclusive to an exclusive interpretation) and not to some other feature. Thus, one would avoid investigating utterances that invite an exclusive understanding of the situation described rather than of the description itself. In example (6) ("Jane is in Paris or in Madrid") above, the exclusive understanding is based on our knowledge that a person cannot be in two places at the same time and need not involve any pragmatic enrichment of the meaning of the word "or". In devising experimental material, it thus becomes important to invent examples where an enriched interpretation is not imposed

[1] From now on, for ease of exposition, we will use the term "scalar" without quotes to refer to the phenomena so described in the neo-Gricean approach. This use, of course, implies no theoretical commitment on our part.

by extra-pragmatic considerations. One can do this by using either examples where participants knowledge is equally compatible with a literal or an enriched interpretation of a scalar term, or examples where knowledge considerations might bias participants in favour of a literal interpretation: in both cases, if one finds evidence of enrichment, one will be confident that it comes from a pragmatic inference about what the utterance meant, rather than from a mere understanding of how the world is.

Second, one would want a paradigm that allows for two identifiable outcomes so that the presence of an enrichment can be indicated by a unique sort of response while a non-enrichment can be indicated by a different response. This is why most of the experiments on scalars described here involve a scenario that could be described by means of a more informative utterance than the test utterance (uttered by a puppet or some other interlocutor). Imagine for example being shown five boxes each containing a token and then being told, "Some boxes contain a token." If one interprets "some" literally (i.e. as compatible with *all*), one would agree with the statement; if one enriches "some" so as to make it incompatible with *all*, one would have to disagree. In such conditions, a participant's response (agrees or disagrees) is revealing of a particular interpretation.

Third, one wants every assurance that an effect is robust. That is, one wants to see the same result over and over again and across a variety of comparable tasks. When two similar studies (for instance two studies investigating different scalar terms but in an equivalent manner) present comparable outcomes, each strengthens the findings of the other. On the other hand, if two very similar experiments fail to produce the same general effects, something is wrong. This does not mean that negative results are necessarily fatal for an experimental paradigm . If one carefully modifies an experiment and it prompts a different sort of outcome than previous ones (and in a predictable manner), it helps determine the factors that underlie an effect. This occurs with the developmental findings to be described below, which have generally shown that children are more likely than adults to *agree* with a weak statement (for instance the statement "Some horses jumped over a fence") when a stronger one would be pragmatically justified (because, in fact, all the horses jumped over a fence). All sorts of follow-up studies have aimed to put this effect to the test. In general, the effect has been resilient; a few studies, however, show that one can get children to appear more adult-like through specific sorts of modifications. For example, experimenters have aimed to verify the effect under conditions where participants are given training or where scenarios are modified to highlight the contrast between the weak utterance and the stronger

scenario. The net result is that the outcomes of these tests collectively help identify the factors that can encourage scalar inference-making.

Fourth, it is important for any experiment to include as many reasonable controls as possible. These are test questions that are similar to the main items of interest, but aim basically to confirm that there is nothing bizarre in the task. For example, if one finds that participants' responses indicate that they enrich "some" but also that the same participants endorse the use of the word "some" to describe a scene where "none" would be appropriate, then there is something questionable about the experiment. This rarely happens (the above example is presented for illustrative purposes only), but one needs to provide assurances to oneself and to readers that such bizarreness can be ruled out. Any decent task will include several controls that lead to uncontroversial responses in order to, in effect, contextualize the critical findings. The studies we will discuss exemplify the four methodological considerations we have just discussed.

Developmental studies

The experimental study of scalar inferences started within the framework of developmental studies on reasoning. Noveck (2001) investigated the way children responded (by agreeing or disagreeing) to a puppet who presented several statements, including one that could ultimately lead to a pragmatic enrichment. All statements, even those that served as controls to confirm that the participants understood the task, were about the contents of a covered box and were presented by a puppet (handled by the experimenter). Participants were told that the contents of the covered box resembled those of one or the other of two other boxes both which were open and with their contents in full view. One open box contained a parrot and the other contained a parrot and a bear. The participants then heard the puppet say[2]:

(17) A friend of mine gave me this (covered) box and said, "All I know is that whatever is inside this box (the covered one) looks like what is inside this box (the one with a parrot and bear) or what is inside this box (the one with just a parrot)."

The participant's task was to say whether or not he agreed with further statements of the puppet. The key item was ultimately the puppet's "underinformative" statement:

[2] The contents of parentheses were not said, but indicated.

(18) There might be a parrot in the box.

Given that the covered box *necessarily* contained a parrot, the statement in (18) can be answered in one of two ways. The participant can "agree" if she interprets "might" literally (so that ...*might*… is compatible with …*must*…) or she can "disagree" if she interprets *might* in an enriched way (where ...*might.*.. is incompatible with …*must.*…). Adults tended to be equivocal with respect to these two interpretations (35% agreed with the statement) while children (5-, 7- and 9-year-olds) tended to interpret this statement in a minimal way, i.e. literally. Collectively, 74% of the children responded by agreeing with the statement in (18). However, not all children were alike.

The five year olds agreed with (18) at a rate of 72% (a percentage that is unlikely to occur by chance, which would yield 50% in such agree/disagree contexts). Nevertheless, they failed to answer many control questions at such convincing rates. For example, when asked to agree or disagree with statements about the bear ("There has to be a bear," "There might be a bear," "There does not have to be a bear," "There cannot be a bear") they answered at levels that were comparable to those predicted by chance (55% correct across the four questions). Seven-year-olds, on the other hand, did manage to answer practically all seven control problems at rates that indicated they understood the task overall (77%). This is why Noveck (2001) reported that seven-year-olds were the youngest to demonstrate competence with this task while at the same time revealing that they preferred the literal interpretation of "might" (at a rate, 80%, that is statistically distinguishable from expectations based on chance). The seven-year-olds thus provided the strongest evidence showing that those linguistically competent children who performed well on the task overall still interpreted "might" in an unenriched way. As one might expect, the nine-year-olds also answered control problems satisfactorily. Response rates indicating unenriched interpretations of "might" were high (69%) and much higher than the adults' but nevertheless were statistically indistinguishable from predictions based on chance suggesting that these children were *beginning* to appear adult-like with respect to (18). Overall, these results were rather surprising for a reasoning study because they indicated that children were more likely than adults to produce a logically correct evaluation of the underinformative modal statement. This sort of response is surprising and rare, but thanks to a pragmatic analysis—where pragmatic enriched interpretations are viewed as likely to result from a richer inferential process than minimal

interpretations that add nothing to semantic decoding—these results had a ready interpretation.

Despite taking every precaution (having numerous control items and sampling many children), one can never exclude that such effects might be the result of some subtle factor beyond the experimenter's intention or control. That is why—especially when encountering counterintuitive results like these—it pays to do follow-ups. There have been essentially two sorts.

The first sort aims to verify the effect. In one follow-up (Noveck, 2001, Exp. 2), the same task as the one above was given to 5-year-olds and 7-year-olds as well as adults, but all participants were given more thorough training to ensure that they understood the parameters of the task. This was done through training on an identical scenario (one box containing a horse and a fish and another just a horse) where pointed questions were asked about the covered box (e.g. *Could there be a fish by itself in the box?*). Overall, the training increased rates of minimal interpretations of "might" across all three ages when it came to the task of Experiment 1. Agreement with a statement like the one in (18) was now 81% for five-year-olds, 94% for seven-year-olds, and 75% for adults. Although rates of such minimal interpretations were statistically comparable across ages, one finds the same trends as in the first Experiment. Seven-year-olds again demonstrated (through performance with the control problems) that they were the youngest to demonstrate overall competence with the task while *tending* to be more likely than adults in retaining a literal interpretation of the weak scalar term. The data also revealed that the extra training encourages adults to behave more "logically" (to stick to the literal meaning of "might"), like the children.

In an effort to establish the developmental effect's reliability and robustness, Noveck (2001, Exp. 3) took advantage of an older study that (a) unintentionally investigated weak scalar expressions among 4- to 7-year-old children and that (b) also failed to show evidence of pragmatic enrichment. Smith (1980) presented statements such as "Some giraffes have long necks" to children and reported that it was surprising to find the children accepting these as true. In a third experiment, therefore, Noveck (2001) essentially continued from where Smith left off. The experiment adopted the same technique as Smith (which included pragmatically felicitous statements such as "Some birds live in cages" as well as statements with "all") in order to verify that the developmental findings of the first two experiments were not flukes. The only differences in this third experiment were that the children were slightly older (8- and 10-years-old) than in the first two studies and that the experimenter was as

"blind" to the intention of the study as the participants (the student who served as experimenter thought that unusual control items such as "Some crows have radios" or "All birds have telephones" were the items of interest). The results showed that roughly 87% of children accepted statements like "Some giraffes have long necks" whereas only 41% of adults did. Again, adults were more likely than children to enrich the interpretation of the underinformative statements (understanding …*some*… as excluding …*all*…) and thus tended to reject them (since all giraffes have long necks). All participants answered the five sorts of control items (25 items altogether) as one would expect.

These data prompted Noveck (2001) to revisit other classic studies that serendipitously contained similar scenarios (ones where a stronger statement would be appropriate but a weaker one is made) to determine whether they tell the same story as "might" and "some". In fact, three studies concerning "or" (Paris, 1973, Braine and Rumain, 1981, Sternberg, 1983), where a conjunctive situation is described with a weaker disjunction, provides further confirming evidence. The authors of these studies also reported counter-intuitive findings showing younger children being, in effect, more logical than adults (children tend to treat "or" inclusively more often than adults). None of these authors, lacking a proper pragmatic perspective, knew how to make sense of these data at the time. All told, this effect appeared robust.

Other follow-up studies have actually taken issue with Noveck's *interpretation* of the findings. In fact, Noveck (2001, p.184) insisted that his data show that children are ultimately less likely than adults to pragmatically enrich underinformative items across tasks; this did not amount to a claim that children lacked pragmatic competence. Still, much work has been aimed at showing that young children are more competent than it might appear. These studies usually take issue with Noveck's Experiment 3 (the one borrowed from Smith, 1980) because it concerns the quantifier "some" (which is of more general interest than "might") and because the items used in that task are admittedly unusual (see Papafragou and Musolino, 2003; Chierchia, Guasti, Gualmini, Meroni, Crain & Foppolo, 2004; Guasti, Chierchia, Crain, Foppolo, Gualmini and Meroni, 2005; Feeney, Scrafton, Duckworth, & Handley, 2005).

We highlight here the main advances of these studies. In two sets of studies, Papafragou and colleagues (Papafragou and Musolino, 2003; Papafragou and Tantalou, 2004) aimed to show that children as young as five are generally able to produce implicatures if the circumstances are right. Actually, Papafragou and Musolino (2003, Experiment 1) first confirmed the developmental effect summarized above by showing that 5-year-olds are less

likely than adults to produce enrichments with "some", "start", and "three", in cases where a stronger term was called for (namely, "all", "finished", and a "larger number," respectively). They then modified the experimental setup in two ways in order to prepare their second experiment. First, before they were tested, participants received training aimed at enhancing their awareness to pragmatic anomalies. Specifically, children were told that the puppet would say "silly things" and that the point of the game was to help the puppet say it better (e.g. they would be asked whether a puppet described a dog appropriately by saying "this is a little animal with 4 legs"). In the event that the child did not correct the puppet, the experimenter did. Second, the paradigm put the focal point on a protagonist's performance. Unlike in their Experiment 1, where participants were asked to evaluate a quantified statement like "Some horses jumped over the fence" (when in fact all the horses did), the paradigm in Experiment 2 raises the expectations that the stronger statement (with "all") might be true. Participants would hear a test statement like, "Mickey put some of the hoops around the pole" (after having his been shown to succeed with all of the hoops), and they were also told how Mickey claims to be especially good at this game and that this is why another character challenges him to get all three around the pole. With these changes, 5-year-olds were more likely to produce enrichments than they were in the first experiment. Nevertheless, the five year olds, even in the second experiment, still produced enrichments less often than did adults. This indicates that – even with training and with a focus on a stronger contrast – pragmatic enrichments require effortful processing among children. [3]

Guasti et al. (2005) argue that pragmatic enrichments ought to be as common among five year olds as they are among adults and further investigated the findings of Noveck (2001) and Papafragou and Musolino (2003). In their first experiment, they replicated the finding of Noveck (2001, Experiment 3) with "some" with 7-year-olds and used this as a baseline to study independently the role of the two factors manipulated by Papafragou and Musolino (2003). One factor was the role of training and how it affects children's proficiency at computing implicatures (Experiments 2 and 3) and the other was the role of placing emphasis on the outcome of a scalar implicature (Experiment 4). Their Experiments 1 through 3 showed that training young participants to give the most specific description of a given

[3] Papafragou and Tantalou (2004) aim to show that five-year-olds can be encouraged to produce scalar inferences and at adult levels. However, we do not discuss their results here because their data are based on a non-standard paradigm in which participants are given no justifiable reason to accept the 'minimal' interpretation of a term such as 'some'. In other words, the paradigm does not provide participants with two clear options. Moreover, much of the study's claims are based on children's self-reports and even these lead to the conclusion that at most 56% of Papafragou and Tantalou's participants derived scalar inferences.

situation can indeed have a major effect on performance. While their initial experiment showed that 7-year-olds accept statements such as “Some giraffes have long necks” 88% of the time (against 50% for adults), when trained in this manner their acceptance rate drops to 52%, becoming adult-like. Nonetheless, this effect is short lasting, i.e. it does not persist when the same participants are tested a week later (Experiment 3). In the last experiment, the authors rendered the *all* alternative more salient in context. This was achieved, for instance, by presenting participants with a story where several characters have to decide whether the best way to go collect a treasure was to drive a motorbike or ride a horse. After some discussion, all of them choose to ride a horse. In this way it is made clearer that the statement subjects have to judge, “some of the characters chose to ride horse,” is underinformative. The results indicated that children are more likely to infer an enriched interpretation in an adult-like manner when the context makes this enrichment highly relevant.

This last finding shows that one can create situations that encourage children to pragmatically enrich weak-sounding statements and to do so in an adult-like manner. It does not alter the fact that in less elaborate scenarios where cues to enrichment are less abundant, seven-year olds do not behave in this manner and it does not tell us what younger children do. Overall, the developmental effect shows that pragmatic enrichments are somewhat effortful. In experimental settings, the required effort can be somewhat lowered or the motivation to perform it may be heightened, but in the absence of such contextual encouragements, younger children faced with a weak scalar term are more likely to stick with its linguistically encoded meaning.

If children had been found to perform scalar inferences by default, this would have been strong evidence in favour of the GCI theory approach. However, taken together, developmental data suggest that, for children, enriched interpretations of scalar terms are not default interpretations. This data is not knock down evidence against GCI theory, because it is compatible with two hypotheses: 1) scalar inferences are not default interpretations for adults either (even if adults are more likely to derive them because they can do so with relatively less effort and because they are more inclined to invest effort in the interpretation of an utterance given their greater ability to derive from it cognitive effects). Or, 2) in the course of development, children become capable and disposed to perform scalar inferences by default. The first hypothesis is consistent with the relevance theory approach while the second is consistent with the GCI approach. To find out which approach has more support, further work had to be done with adults.

Time course of comprehension among adults

As we mentioned before, GCI theory implies that a literal interpretation of a scalar, resulting from the cancellation of default enrichment, should take longer than an enriched interpretation, whereas relevance theory, denying that enrichment takes place by default, implies that an enriched interpretation, being computed when needed to meet contextual expectations of relevance, should take longer than a literal one. What is needed to test these contrasting predictions are experiments manipulating and measuring the time course of the interpretation of statements with weak scalar terms.

As in the developmental tasks, one wants to make sure that enriched interpretations are clearly identifiable through specific responses, that the tasks used includes a variety of controls, and that the effect is reliable and robust. One way to identify enriched vs. literal interpretations is provided by earlier studies where participants were asked to judge true or false statements (such as "some elephants are mammals") that could either be construed as literally true but underinformative, or in an enriched manner (as implying …*not all*…) and false. Hence participants' truth-value judgements reflect their literal or enriched interpretation.

As we indicated, prior work was critical to developing the appropriate measures. In fact, Rips (1975) unintentionally included the right sort of cases when looking at other issues of categorization and with materials such as "some congressmen are politicians." He examined the effect of the interpretation of the quantifier by running two studies, one in which participants were asked to treat "some" as meaning *some and possibly all* and another where they were asked to treat "some" as meaning *some but not all*. This comparison demonstrated that participants given the *some but not all* instructions in one experiment responded more slowly than those given the *some and possibly all* instructions in another. Despite these indications, Rips modestly hedged when he concluded that "of the two meanings of *Some*, the informal meaning *may* be the more difficult to compute" (italics added). To make sure that Rips's data were indeed indicative of a slowdown related to *Some but not all* readings, Bott & Noveck (2004) ran a series of four experiments that followed up on Rips (1975) and essentially verified that enriched interpretations take longer than literal ones.

Bott and Noveck's categorization task involved the use of underinformative items (e.g. "Some cows are mammals") and five controls that varied the quantifier (*Some* and *All*), the

category-subcategory order, as well as proper membership. The 6 types of statements are illustrated with the six possible ways one can employ the subcategory *elephants* below, but it should be pointed out that the paradigm was set up so that the computer randomly paired a given subcategory with a given category while verifying that, at the end of each experimental session, there were nine instances of each type:

(19) (a) Some elephants are mammals (Underinformative).
(b) Some mammals are elephants.
(c) Some elephants are insects
(d) All elephants are mammals.
(e) All mammals are elephants.
(f) All elephants are insects.

In the first Experiment, a sample of 22 participants was presented with the same task twice, once with instructions to treat "some" as meaning *Some and possibly all* and once with instructions to treat "some" as meaning *Some but not all* (and, of course, the order of presentation was varied). When participants were under instruction, in effect, to engage the scalar inference, they were shown to be less accurate and take significantly longer to respond to the underinformative items (like those in (19a)). Specifically, when instructions called for a *Some but not all* interpretation, rates of correct responses to the Underinformative item (i.e. judging the statement "false") were roughly 60%; when instructions called for a *Some and possibly all* interpretation rates of correct responses to the Underinformative item (i.e. judging the statement "true") were roughly 90%. For the control items, rates of correct responses were always above 80% and sometimes above 90%. One can see that the Underinformative case in the *Some but not all* condition provides exceptional data.

The reaction time data showed that the correct responses to the Underinformative item in the *Some but not all* condition were exceptionally slow. It took roughly 1.4 seconds to correctly evaluate the underinformative statements in the *Some but not all* condition and around .8 seconds in the *Some and possibly all* conditions. To answer the control items—across both sorts of instructions—took at most 1.1 seconds but more often around .8 to .9 seconds. Thus, the underinformative statement in the *Some but not all* condition is the one most affected by the instructions. All this confirms Rips's initial findings. More importantly, there is not a single indication that interpreting "some" to mean *some but not all* is an

effortless or quasi effortless step. Again, a default view of scalar inference would predict that under *Some but not all* instructions, responses to Underinformative statements would require less time than responses under *Some and possibly all* instructions. According to an account based on relevance theory, one should find the opposite. The data more readily support the relevance-theoretic account.

A potential criticism of this Experiment is that the lower accuracy and the slowdown might be due to a response bias in favour of positive rather than negative response, given that the correct response to the Underinformative statement with the *Some and possibly all* instructions is to say "True" while the correct response to the Underinformative statement with the *Some but not all* instructions is to say "False." To allay concerns regarding such a potential response bias, Bott and Noveck demonstrated experimentally that the effects linked to pragmatic effort are not simply due to hitting the "False" key.

In a second experiment, the paradigm was modified so that the same overt response could be compared across both sorts of instructions; this way, participants' response choice (True vs. False) could not explain the observed effects. In order to arrive at this comparison, participants were not asked to agree or disagree with first-order statements such as those in (19), but with second order statements made about these first-order statements. For example, participants were presented with the two statements: "Mary says the following sentence is false" / "Some elephants are mammals." They were then asked to agree or disagree with Mary's second-order statement. In such a case, participants instructed to treat "some" as meaning *Some but not all* should agree, whereas participants instructed to treat "some" as meaning *Some and possibly all* should disagree, reversing the pattern of positive and negative response of the previous experiment.

The results from this second experiment were nevertheless remarkably similar to those of the first one. Here, when participants were under instruction to, in effect, draw the scalar inference, they were less accurate and took significantly longer to respond correctly to the underinformative item. When "agree" was linked with instructions for a *Some but not all* interpretation, rates of correct responses were roughly 70%; when "agree" was linked with instructions for a *Some and possibly all* interpretation, rates of correct responses were roughly 90%. For all control items, rates of correct responses were always above 85% and often above 90%. One can see that, once again, the Underinformative case in the *Some but not all* condition provides exceptional data. The reaction time data also showed that the correct "agree" responses to the Underinformative item in the *Some but not all* condition were

exceptionally slow. It took nearly 6 seconds to evaluate the underinformative statements correctly when "agree" was linked with instructions for a *Some but not all* interpretation and around 4 seconds when "agree" was linked with instructions in the *Some and possibly all* condition (all reaction times were longer than in the previous experiment due to the *Mary says* statement). The control items across both sorts of instructions took on average around 4.5 seconds and never more than 5 seconds. Again, the experiment demonstrated that any response that requires a pragmatic enrichment implies extra effort.

Both of these experiments, though inspired by previous work, are arguably unnatural. It is unusual to instruct participants in a conversation, as was done in Experiment 1, as to how they should interpret the word "some"; the second experiment doubles the complexity by compelling participants to make metalinguistic judgments from statements like *Mary says the following is false*. Bott and Noveck's third experiment simplified matters by asking participants to make true/false judgments about the categorical statements themselves and without prior instruction. With this sort of presentation, there is no useful sense in which a response is "correct" or not. Rather, responses reveal the participants literal or enriched interpretation and can be compared in terms of reaction time.

Roughly 40% of participants responded "true" to Underinformative items and 60% "false". This corresponds to the rates found among adults in Noveck's developmental studies (also see Noveck & Posada, 2003; Guasti et al, 2005). The main finding was that mean reaction times were longer when participants responded "false" to the underinformative statements than when they responded "true" (3.3 seconds versus 2.7, respectively). Furthermore, "false" responses to the underinformative statements appear to be slower than responses to all of the control statements (including three, (18c), (18e), and (18f), that require a "false" response). The "true" response was made at a speed that was comparable to all of the control items.

In their last experiment, Bott and Noveck's varied the time available to participants to respond to the statements. The rationale for this design was that, if as implied by GCI theory, literal interpretations of weak scalar terms take longer than the default enriched interpretations, then limiting the time available should decrease the rate of literal interpretations and increase the rate of enriched ones. On the other hand, if as implied by relevance theory, enriched interpretations take longer, then limiting the time should have the opposite effect. While following the same general procedure as the prior experiments (asking participants to judge the veracity of categorical statements), the paradigm manipulated the

time available for the response. In one condition, participants had a relatively long time to respond (3 seconds), while in the other they had a relatively short time to respond (.9 seconds). Only the time to *respond* was manipulated. To control for uptake, participants were presented with the text one word at a time and at the same rate in both conditions, thus there is no possibility that participants in the Short-lag condition spend less time reading the statements than those in the Long-lag condition.

Bott and Noveck reported that when a shorter period of time was available for participants to respond, they were more likely to respond "True" to Underinformative statements (indicating a literal interpretation). 72% of participants responded true in the "Short-lag" condition and 56% did so in the "Long-lag" condition. This strongly implies that they were less likely to derive the scalar inference when they were under time pressure than when they were relatively pressure-free. As in all the prior experiments, control statements provide a context in which to appreciate the differences found among Underinformative statements. These showed that performance among control statements in the Short Lag condition was quite good overall (rates of correct responses ranged from 75%-88%) and that, as one would expect, rates of correct performance among the control items *increased* with added time (by 5% on average). The contrast between a percentage that drops with extra time (as is the case for the Underinformative statements) and percentages that increase provide a unique sort of interaction confirming that time is necessary to provoke scalar inferences.

The experiments we have described so far take into account the four methodological considerations we discussed earlier and allow well-controlled measure of a dependent variable: the rate or the speed of literal vs. enriched interpretations of weak scalar terms. Together, they provide strong evidence that an enriched interpretation of a weak scalar term requires more processing time than an unenriched, literal interpretation, as predicted by relevance theory and contrary to the prediction implied by GCI theory.

Still, one might argue that the categorization tasks used, even if methodologically sound from an experimental psychology point of view, are too artificial to test pragmatic hypotheses. If the argument were that laboratory tasks are somehow irrelevant to pragmatics, we would argue that the onus of the proof is on the critics: after all, participants bring to bear on experimental verbal tasks their ordinary pragmatic abilities, just as they do in any uncommon form of verbal exchange. In particular, if it is part of adult pragmatic competence to make scalar inferences by default, it would take some arguing to make it plausible that an experimental setting somehow inhibits this basic disposition. On the other hand, if the

argument is that fairly artificial laboratory experiments are not enough and that they should be complemented with more ecologically valid designs, we agree. Happily, Breheny, Katsos and Williams (2006) have provided just this kind of welcome complement.

Following up on a procedure from Bezuidenhout & Cutting (2004), Breheny et al. presented disjunctive phrases (such as "the class notes or the summary") in two kinds of contexts: Lower-bound contexts (where the literal reading of a scalar term is more appropriate as in (20) below), and Upper-Bound contexts (where the enriched reading of the scalar is more appropriate as in (21) below). These were presented as part of short vignettes (along with many "filler" items to conceal the purpose of the study) and participants' reading times were measured. More specifically, participants were asked to read on a computer screen short texts that were presented one fragment at a time, and to advance in their reading by hitting the space bar (the slashes in (20) and (21) delimit fragments).

(20) *Lower-bound context*

John heard that / the textbook for Geophysics / was very advanced. / Nobody understood it properly./ He heard that / if he wanted to pass the course / he should read / *the class notes or the summary*.

(21) *Upper-bound context*

John was taking a university course / and working at the same time. / For the exams / he had to study / from short and comprehensive sources./ Depending on the course, / he decided to read / *the class notes or the summary*.

If, in such a task, one found shorter reading times in the Upper-bound contexts that call for scalar inferences than in the Lower-bound contexts where the literal interpretation is more appropriate, this would support the GCI claim that scalar inferences are made by default. Findings in the opposite direction would support the relevance theory account. What Breheny et al. found is that phrases like the *class notes or the summary* took significantly longer to process in Upper-bound contexts than in Lower-bound contexts, a result consistent with findings reported above.

Conclusions

The experimental work we have summarized here verifies predictions derived from relevance theory, and falsifies predictions derived from GCI theory. Does this mean that relevance theory is true and GCI theory is false? Of course not. Nevertheless, these results should present a serious problem for GCI theorists. It is quite possible however that they will find a creative solution to the problem. They might for instance show that, in spite of the methodological precautions we have outlined, the studies reported failed to eliminate some uncontrolled factor, and that better studies provide evidence pointing in the opposite direction. They might, more plausibly, revise their theory so as to accommodate these results. One line of revision would be to reconsider the idea that GCI are default inferences (or to water down the notion of default to the point where it does not anymore have implications for processing time). After all, not all neo-Griceans agree with Levinson's account of GCI (see in particular Horn 2004, 2006). Still, it is worth noting that, if scalar inferences are not truly default inferences and involve each and every time paying attention to what the speaker chose not to say, then we are back to the worry that such inferences are excessively cumbersome. Generally speaking, experimental findings such as those we have summarised here should encourage neo-Griceans to work out precise and plausible implications of their approach at the level of cognitive processing.

Relevance theorists are not challenged in the same way by the work we have described—after all, their prediction is confirmed—, but they should be aware that this prediction could be made from quite different theoretical points of view: it follows from relevance theory, but relevance theory does not follow from it. They might then try to develop aspects of these experiments that could give positive support to more specific aspects of the theory. For instance, according to the theory, hearers aim at an interpretation that satisfies their expectations of relevance and the relevance of an interpretation varies inversely with the effort needed to derive it. It should then be possible to cause participants to choose a more or a less parsimonious interpretation by increasing or decreasing the cognitive resources available to participants for the process of interpretation. The fourth experiment of Bott and Noveck (2004) can be seen as a first suggestive step in this direction.[4]

As we have just explained, we do not expect readers to form a final judgement on the respective merits of GCI theory and Relevance theory on the basis of the experimental evidence presented. What we do hope is to have convinced you that, alongside other kinds of

[4] For other experimental explorations based on relevance theory, see Van der Henst & Sperber 2004.

data, properly devised experimental evidence can be highly pertinent to the discussion of pragmatic issues, and that pragmatists—and in particular students of pragmatics—might greatly benefit from becoming familiar with relevant experimental work and from contributing to it (possibly in interdisciplinary ventures).

References

Bezuidenhout, A. & Cutting, J.C. (2002). Literal meaning, minimal propositions, and pragmatic processing. Journal of Pragmatics, 34, 433-456.

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language, 51,* 437-457.

Braine, M. & Rumain, B. (1981). Children's comprehension of "or": Evidence for a sequence of competencies. *Journal of Experimental Child Psychology, 31,* 46-70.Breheny, R., Katsos, N., & Williams, J. (2006). Are scalar implicatures generated on-line by default? *Cognition, 100 (3),* 434-463.

Breheny, R., Katsos, N., & Williams, J. (2006). Are generalized scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition 100(3),* 434-463.

Chierchia, G., Guasti, T., Gualmini, A., Meroni, L., Crain, S. & Foppolo, F. (2004). Adults and children's semantic and pragmatic competence in interaction. In *Experimental Pragmatics*, I. A. Noveck & D. Sperber (Eds.). Basingstoke: Palgrave Macmillan.

Feeney, A., Scrafton, S., Duckworth, A. and Handley, S. J. 2004: The Story of *Some*: Everyday Pragmatic Inference by Children and Adults. *Canadian Journal of Experimental Psychology, 58,* 2, 121-132.

Guasti, M. T., Chierchia, G.,Crain, S., Foppolo, F., Gualmini, A. & Meroni. L. (2005). Why children and adults sometimes (but not always) compute implicatures. *Language and Cognitive Processes*, 20 (5), 667–696

Horn, L. R. (1972). *On the Semantic Properties of Logical Operators in English.* Ph. D. Dissertation, UCLA.

Horn, L. R. (2004). Implicature. In Horn and Ward (eds.), 3-28.

Horn, L. R. (2006). The Border Wars: A neo-Gricean perspective. In Turner and von Heusinger (eds.).

Levinson, S. (2000). *Presumptive Meanings*. MIT Press.

Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition, 78*(2), 165-188.

Noveck, I. A. (2004). Pragmatic inferences related to logical terms. In I. A. Noveck & D. Sperber (Eds.), *Experimental Pragmatics*. Basingstoke: Palgrave.

Noveck, I. A. & Posada, A. (2003). Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language, 85, 203-210.*

Noveck, I. A. & Sperber, D. (2004), *Experimental Pragmatics*. Basingstoke: Palgrave.

Papafragou, A., & Musolino, J. (2003). Scalar implicatures: experiments at the semantics-pragmatics interface. *Cognition, 86*(3), 253-282.

Papafragou, A; & Tantalou, N. (2004). Children's computation of implicatures. *Language Acquisition 12(1),* 71-82.

Paris, S. G. (1973). Comprehension of language connectives and propositional logical relationships. *Journal of Experimental Child Psychology, 16(2),* 278-291.

Rips, L. J. (1975). Quantification and semantic memory. *Cognitive Psychology, 7*(3), 307-340.

Smith, C. L. (1980). Quantifiers and question answering in young children. *Journal of Experimental Child Psychology, 30(2),* 191-205.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and Cognition*. Oxford: Basil Blackwell.

Sternberg, R. J. (1979). Developmental patterns in the encoding and combination of logical connectives. *Journal of Experimental Child Psychology, 28*(3), 469-498.

Van der Henst, J.B., Sperber, D. (2004). Some experimentally testable implications of relevance theory. In I. A. Noveck & D. Sperber (Eds.), *Experimental Pragmatics*. Basingstoke: Palgrave.

Wilson, D. & Sperber, D., (2004). Relevance Theory, in G. Ward et L. Horn (Eds.), *Handbook of Pragmatics*. Blackwell.