

Some utterances are underinformative: The onset and time course of scalar inferences

Lewis Bott & Ira A. Noveck
Institut des Sciences Cognitives
Centre National de la Recherche Scientifique
Lyon, France

RUNNING HEAD: Time course of scalars

Address for correspondence:

Lewis Bott
Rm 873, Department of Psychology,
New York University,
6 Washington Place,
New York, NY 10003,
USA

Lewis.Bott@nyu.edu

tel. : +1-212-998-7902

fax.: +1-212-998-7847

Abstract

When Tarzan asks Jane *Do you like my friends?* and Jane answers *Some of them*, her underinformative reply implicates *Not all of them*. This *scalar inference* arises when a less-than-maximally-informative utterance implies the denial of a more informative proposition. Default Inference accounts (e.g. Levinson, 1983; 2000) argue that this inference is linked to lexical items (e.g. *some*) and is generated automatically and largely independently of context. Alternatively, Relevance Theory (Sperber and Wilson, 1986/1995) treats such inferences as contextual and as arriving effortfully with deeper processing of utterances. We compare these accounts in four experiments that employ a sentence verification paradigm. We focus on underinformative sentences, such as *Some elephants are mammals*, because these are false with a scalar inference and true without it. Experiment 1 shows that participants are less accurate and take significantly longer to answer correctly when instructions call for a *Some but not all* interpretation rather than a *Some and possibly all* interpretation. Experiment 2, which modified the paradigm of Experiment 1 so that correct responses to both interpretations resulted in the same overt response, reports results that confirm those of the first Experiment. Experiment 3, which imposed no interpretations, reveals that those who employed a *Some but not all* reading to the underinformative items took longest to respond. Experiment 4 shows that the rate of scalar inferences increased as a permitted response time did. These results argue against a neo-Gricean account and in favor of Relevance theory.

There is a growing body of psycholinguistic work that focuses on the comprehension of logical terms. These studies can be broken down into two sets. One investigates the way logical inferences are made on-line in the context of story comprehension (e.g. Lea, O'Brien, Fisch, Noveck, & et al., 1990; Lea, 1995) . In this approach, the comprehension of a term like *or* is considered to be tantamount to knowing logical inference schemas attached to it. For *or* it would be or-elimination (where the two premises - $p \text{ or } q$; $\text{not-}q$ – imply p). The other line of research investigates non-standard existential quantifiers, such as *few* or *a few*, demonstrating how the meanings of quantifiers - besides conveying notions about amount - transmit information about the speaker's prior expectations as well as indicate where the addressee ought to place her focus (Moxey, Sanford, & Dawydiak, 2001; Paterson, Sanford, Moxey, & Dawydiak, 1998, Sanford, Moxey, & Paterson, 1996). For example, positive quantifiers like *a few*, put the focus on the quantified objects (e.g. those who got to the match in *A few of the fans went to the match*) while negative quantifiers like *few* place the focus on the quantified objects' complement (e.g. those fans who did not go to the match in *Few of the fans went to the match*).

In the present paper, we investigate a class of inference - which we will refer to as a *scalar inference* - that is orthogonal to the ones discussed above, but is arguably central to the way listeners treat logical terms. These arise when a less-than-maximally-informative utterance is taken to imply the denial of the more informative proposition (or else to imply a lack of knowledge concerning the more informative one). Consider the following dialogues:

- 1) Peter: Are Cheryl and Tony coming for dinner?
Jill: We are going to have Cheryl or Tony.

2) John: Did you get to meet all of my friends?

Robyn: Some of them.

In (1), Jill's statement can be taken to mean that *not both* Cheryl and Tony are coming for dinner and, in (2), that Robyn did not meet all of John's friends. These interpretations are the result of scalar inferences, which we will describe in detail below. Before we do so, note that the responses in each case are compatible with the questioner's stronger expectation from a strictly logical point of view; if Jill knows that both Cheryl and Tony are coming, her reply is still true and if in fact Robyn did meet all of John's friends, she also spoke truthfully. *Or* is logically compatible with *and* and *some* is logically compatible with *all*.

Linguistic background

Scalar inferences are examples of what Paul Grice (1989) called *generalized implicatures* as he aimed to reconcile logical terms with their non-logical meanings. Grice, who was especially concerned by propositional connectives, focused on logical terms that become, through conversational contexts, part of the speaker's overall meaning. In one prime example, he described how the disjunction *or* has a weak sense, which is compatible with formal logic's \vee (the inclusive-or), but as benefiting from a stronger sense (*but not both*) through conversational uses (making the disjunction exclusive). What the disjunction *says*, he argued, is compatible with the weaker sense, but through conversational principles it often *means* the stronger one. Any modern account of the way logical terms are understood in context would not be complete without considering these pragmatic inferences.

Grice's generalized implicatures were assumed to occur very systematically although the context may be such that they do not occur. These were contrasted with particularized implicatures, which were assumed to be less systematic and always clearly context

dependent. His reasons for making the distinction had to do with his debates with fellow philosophers on the meaning of logical connectives and of quantifiers, and not with the goal of providing a processing model of comprehension, and there is some vagueness in his view of the exact role of the context in the case of generalized implicatures (see Carston, 2002, pages 107-116). In summary, Grice can be said to have inspired work on implicatures (by providing a framework), but there is not enough in the theory to describe, for example, how a scalar inference manifests itself in real time.

Pragmatic theorists, who have followed up on Grice and are keen on describing how scalar inferences actually work, can be divided into two camps. On the one hand, there are those who assume that the inference generally goes through unless subsequently cancelled by the context. That is, scalars operate on the (relatively weak) *terms* - the speaker's choice of a weak term implies the rejection of a stronger term from the same scale. To elucidate with disjunctions, the connectives *or* and *and* may be viewed as part of a scale (<*or*, *and*>), where *and* constitutes the more informative element of the scale (since *p and q* entails *p or q*). In the event that a speaker chooses to utter a disjunctive sentence, *p or q*, the hearer will take it as suggesting that the speaker either has no evidence that a stronger element in the scale, i.e. *p and q*, holds or that she perhaps has evidence that it does not hold. Presuming that the speaker is cooperative and well informed the hearer will tend to infer that it is not the case that *p and q* hold, thereby interpreting the disjunction as exclusive. A strong default approach has been defended by *Neo-Griceans* like Levinson (2000) and to some extent by Horn (1984, page 13). More recently, Chierchia (in press) and colleagues (Chierchia, Guasti, Gualmini, Meroni, and Crain, 2001) have essentially defended the strong default view by making a syntactic distinction with respect to scalar terms: When a scalar is embedded in a downward-entailing context (e.g. negations and question forms), Chierchia and colleagues predict that one would

not find the production of scalar inferences (also see Noveck et al., 2002). Otherwise, Chierchia and colleagues do assume that scalar inferences go through.

For the sake of exposition, we focus on Levinson (2000) because he has provided the most extensive proposal for the way pragmatically enriched “default” or “preferred” meanings of weak scalar terms are put in place. Scalars are considered by Levinson to result from a Q-heuristic, dictating that “What isn’t said isn’t (the case).” It is named *Q* because it is directly related to Grice’s (1989) first maxim of quantity: *Make your utterance as informative as is required*. In other words, this proposal assumes that scalars are general and automatic. When one hears a weak scalar term like *or*, *some*, *might* etc. the default assumption is that the speaker knows that a stronger term from the same scale is not warranted or that she does not have enough information to know whether the stronger term is called for. Default means that relatively weak terms prompt the inference automatically - *or* becomes *not both*, *some* becomes *some but not all* etc. Also, a scalar inference can be cancelled. If this happens, it occurs subsequent to the production of the scalar term.

On the other hand, there are pragmatists who argue against the default view and in favor of a more contextual account. Such an account assumes that an utterance can be inferentially enriched in order to better appreciate the speaker’s intention, but this is not done on specific words as a first step to arrive at a default meaning. We focus on Relevance Theory because it arguably presents the most extensive contextualist view of pragmatic inferences in general and of scalar inferences in particular (see Post face of Sperber and Wilson, 1995). According to this account, a scalar is but one example of pragmatic inferences which arise when a speaker intends and expects a hearer to draw an interpretation of an utterance that is relevant enough. How far the hearer goes in processing an utterance’s meaning is governed by principles concerning effect and effort; namely, listeners try to gain as many effects as possible for the least effort.

A non-enriched interpretation of a scalar term (the one that more closely coincides with the word's meaning) could very well lead to a satisfying interpretation of this term in an utterance. Consider *Some monkeys like bananas*: A weak interpretation of *Some* (with which the utterance can be glossed as *Some and possibly all monkeys like bananas*) can suffice for the hearer and not require further pragmatic enrichment. The potential to derive a scalar inference comes into play when an addressee applies relevance more stringently. A scalar inference could well be drawn by a hearer in an effort to make an utterance, for example, more informative (leading to an utterance that can be glossed as *Some but not all monkeys like bananas*). Common inferences like scalars are inferences that optionally play a role in such enrichment; they are not steadfastly linked to the words that could prompt them. If a scalar does arrive in a context that enriches an underinformative utterance, all things being equal the inference ought to be linked with extra effort.

One can better appreciate the two accounts by taking an arbitrary utterance (3) and comparing the linguistically encoded meaning (4a) and the meaning inferred by way of scalar inference (4b):

- (3) Some X are Y.
- (4) a. Some and possibly all X are Y (logical interpretation).
 b. Some but not all X are Y (pragmatic interpretation).

Note that (4a) is less informative than (4b) because the former is compatible with any one of four possible treatments of *some* in (3). That is, *some X are Y* can be viewed as having 4 representations in order to be true, where i) X is a subset of Y, ii) Y is a subset of X, iii) X and Y overlap, and where iv) X and Y coincide. With interpretation (4b), only (ii) and (iii) remain as possibly true. The interpretation represented by (4b) reduces the range of meanings

of *some*. According to Levinson, the interpretation in (4b) is prepotently adopted through the Q-heuristic. This becomes the default meaning unless something specific in the context leads one to cancel (4b) and to then adopt the reading in (4a).

According to Relevance Theory, a listener starts with the interpretation that corresponds with the meaning of the words, like in (4a); if that reading is satisfactory to the listener, she will adopt it. However, if the listener aims to make (3) more relevant, e.g. more informative, she will adopt (4b) instead. Given that (4b) arrives by way of a supplementary step (scalar inference), there is a cost involved (i.e. cognitive effort). This amounts to deeper processing but at a cost.

We propose that the two explanations can be separated by looking at the time course of processing sentences involving scalar inference. Consider first the neo-Gricean view. This account assumes that the ‘default’ meaning is the *initial* interpretation for the weak scalar term, which includes the negation of the stronger elements on the scale. It follows that to interpret the sentence without the inference, the listener must pass through a stage where the scalar inference has been considered and then rejected on the basis of contextual information. Thus, the time taken to process a sentence without a scalar inference must be greater than or equal to one in which a scalar inference is present. In contrast, Relevance Theory considers the weaker sense of a scalar term to be considered first, and only if it is sufficiently relevant is the inference made to deny the stronger term. Comparing the processing times of sentences that have been interpreted with a scalar inference to those that have been interpreted without the inference can therefore be used as evidence to distinguish the two theories.

We should state at this point that although Levinson (1983, 2000) believes default rules and heuristics are an integral part of his theory and that processing issues are central, his account has not explicitly made the processing predictions that we suggest above.

Nevertheless, we feel that there is some intrinsic interest in generating predictions from such

a default model because this idea is at the heart of many neo-Gricean claims. To avoid confusion between predictions based on a range of neo-Gricean accounts and on the default model we test here, we refer to the processing predictions described above as stemming from a Default Inference (DI) account.

Psychological background

Response time experiments in which the interpretation of a scalar term has been important have generally instructed their participants to interpret the term in a strictly logical way (i.e. without the scalar inference). For example, Meyer (1970) informed participants to treat *some* to mean *some and possibly all* in a sentence verification tasks with sentences like *some pennies are coins*. To our knowledge, the only early psychological study to take an interest in the potentially conflicting interpretations of such underinformative sentences was Rips (1975). Rips investigated how participants make category judgments by using sentence verification tasks with materials like *some congressmen are politicians*. He examined the effect of the quantifier interpretation by running two studies, one in which participants were asked to treat *some* as *some and possibly all* and another where they were asked to treat *some* as *some but not all*. This comparison demonstrated that the 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). His reaction is not uncommon. Many colleagues share the intuition that the pragmatic interpretation seems more natural. In any case, this is an initial finding that goes in favor of the Relevance account.

Surprisingly, this finding has not led to any follow-up experiments. We consider four reasons for this. First, Rips’s (1975) initial investigation was only incidentally concerned

with pragmatic issues so it did not put a spotlight on this very interesting finding. Second, until recently, linguistic-pragmatic issues have not been central to traditional cognitive investigations (see Noveck, 2001). Third, skeptics might point out that Rips's effect relies on data collected across two experiments that were ultimately not comparable. It could be argued that his result may be due to sampling bias because participants were not allocated randomly to the two instructions conditions; also, the experiment that requested a logical interpretation (*some and possibly all*) included five types of sentences whereas the experiment that requested a pragmatic interpretation included four. Finally, a task requiring participants to apply certain kinds of interpretations is arguably artificial and does not necessarily capture what occurs under more natural circumstances.

We now turn to the four experiments in the paper. We investigate responses to underinformative categorical statements like *some elephants are mammals*¹ as we compare the Default Inference and Relevance Theory accounts of scalar inference onset. In Experiment 1, we replicated Rips (1975, experiments 2 and 3) in one overarching procedure to address some of the concerns mentioned earlier. Furthermore, we make comparisons between the underinformative sentences and control sentences that were not made in Rips's original experiment. Experiment 2 uses the same paradigm as Experiment 1 but changes the presentation of the sentences and the response options so that correct responses to the two sentences that make up the most critical comparison require the same response key. Experiment 3 was similar to Experiment 1, except that it did not provide precise instructions about the way one ought to treat *some*. All three of these experiments allow us to make a comparison between the Default Inference account and Relevance Theory. According to the

¹ The experiments were all conducted in French, where we used the word *certain*s as the translation of *some* in English. The distribution of scalar inferences associated with *certain*s is similar to that of *some* in English.

Default Inference model, a response prompting an implicature should be faster than one that requires its cancellation. In contrast, Relevance Theory would predict that the minimal meaning of *some* allows for an immediate treatment of a statement that has no need for an implicature and that the production of the implicature arises when participants apply more effort to treating the weak quantifier. The final experiment is a direct test of Relevance Theory with this paradigm. Participants made the same judgments as in Experiments 1 and 3, but we manipulated the time available for responding. A reduction in the processing time was expected to reduce the possibility of producing the scalar inference.

Experiment 1

Experiment 1 presents categorical sentences and asks participants to provide True / False judgments. Examples of the six types of sentences included are shown in Table 1, translated from the French. Sentences referred to as T1 are the underinformative statements described before. In one experimental session, participants were instructed to interpret the quantifier *some* to mean *some and possibly all*, which we refer to as the Logical condition. In another session, they were told to interpret *some* to mean *some but not all*, which we will refer to as the Pragmatic condition. Central to our interests is the speed of response to the T1 sentences under the two conditions. According to the Default Inference account, correct responses in the Logical condition ought to take longer than those in the Pragmatic condition because a logical interpretation requires one to undo the default inference. If the correct response in the Pragmatic condition takes longer to engage than the Logical one, then that would provide evidence against a Default Inference account.

Insert Table 1 about here

We employed several different types of control sentences to be faithful to Rips's (1975) original design and to ensure that any effects we observed were not due to an artifact of the instructions. These control sentences are shown in Table 1, together with the correct response (true or false) associated with each sentence type. Sentences T2 and T3 are statements containing *some* but which have different category constructions. Sentence T2 is a case in which the category is the subject and the member is in the predicate (*some mammals are monkeys*) and T3 is a case in which category membership is false (*some monkeys are insects*). Sentences T4, T5 and T6 use the quantifier *all* and have equivalent category structures to T1-T3. We expect that if differences observed under the two instruction conditions on the T1 sentences are due primarily to the effects of the inference (and not to a general effect of the instructions), then the difference between the Logical and Pragmatic instruction conditions will be largest on the underinformative sentences. Furthermore, if a Relevance theory account is correct, more time should be required to evaluate T1 sentences under pragmatic instructions than to evaluate the control sentences under pragmatic instructions. This is because the inference is not necessary for the control sentences.

Method

Participants. Twenty-two participants were recruited from the area around Lyon. All were native French speakers and were either unpaid or received a gift worth approximately 5 Euros for participation.

Stimuli and Design. The experiment was split into two sessions; one session required participants to interpret *some* in a pragmatic way and the other in a logical way. Before each session, participants saw appropriate instructions and went through a practice phase which included feedback. The order of Logical and Pragmatic sessions was counterbalanced across participants. The experiment took place entirely in French, although the English translations are presented in the paper.

Participants saw six types of sentences, which are shown in Table 1 together with an example of each. In each experimental session, participants saw 9 examples of each type of sentence, making a total of 54 sentences. For each participant, the experimental sentences were generated randomly from a base of 6 categories and 9 exemplars from each of these categories (see Appendix). For example, a participant might see the sentences: *some trout are fish; some mammals are elephants; some monkeys are insects; all parrots are birds; all insects are mosquitoes; all robins are shellfish*; while a different participant would see a completely different set of sentences. Each exemplar from a category was used once only in the experimental session, so no participant would see both *some monkeys are insects* and *some monkeys are mammals*. This randomization procedure was adopted to eliminate, or at least minimize, any unwanted effects of frequency or typicality on the reaction times.²

Before each experimental session, participants saw 16 practice statements concerning categories not tested in the experimental session, for example trees and clothes. These sentences were of the form T1, T2, T4, or T5. We used only four types of sentences because we wished to be consistent with Rips (1975) and because we felt the nature of sentence types T3 and T6 were obvious and therefore needed no training. Participants also saw 5 dummy sentences at the beginning of the experiment to avoid problems associated with starting the experimental phase. All participants saw exactly the same practice and dummy sentences.

Participants made their response using the computer keyboard and they were given feedback on all trials, consisting of the word ‘correct’ or ‘incorrect’ appropriately. Of course, the feedback remained the same across conditions for sentences T2-T6. For T1, however, the

² One will note that the Appendix includes *spider* as an exemplar of insect, although it is an arachnid. Our analyses indicate that this does not pose a problem for participants in our experiments, who treat spiders like other exemplars from the insect category.

feedback for the correct response was provided as a function of the type of instructions received.

The procedure used for practice trials was identical to the experimental trials. However, participants were encouraged to ask questions during the practice phase and to work independently during the experimental session. Participants were not told of the existence of the dummy sentences.

Procedure. Participants were presented with instructions at the beginning of each experimental session. During the first session, participants were not made aware that they would be doing a second session afterwards. The relevant instructions for the Pragmatic condition were as follows (translated from French): “Half of the sentences start with the word *some*, like *some daffodils are flowers*. This word, *some*, can be understood in several ways. We would like you to understand it as *some but not all*. Thus, a sentence like *some daffodils are flowers* should be considered false because, in fact, all daffodils are flowers.” For the Logical condition, the last two sentences of these instructions were changed to : “We would like you to understand it as *some and possibly all*. Thus, a sentence like *some daffodils are flowers* should be considered true, even though we know that all daffodils are flowers.”³

After the completion of one experimental session, participants were presented with a second set of instructions, this time asking them to treat *some* differently (if they received pragmatic instructions in the first session, they were given instructions to respond logically in

³ An anonymous reviewer asked whether the logical instructions -- to treat *some* as *some and possibly all* -- might lead to confusion for T2 sentences because it leads to, e.g., “...possibly all mammals are elephants”. We point out that our choice of words for the logical interpretation is one of several that could convey a minimal meaning for *Some* (e.g., consider instead “at least one”), but it is best suited to be compared to the pragmatic interpretation, which necessarily includes the word “all”. In any case, participants were also given an example sentence during the general instructions to clarify the intended meaning, and a training session (with feedback) to remove any such ambiguity. Furthermore, there is nothing from the data suggesting that

the second and vice versa). There was then another practice phase, followed by the second experimental phase. The stimuli in the practice phase remained the same as in the first session, while those in the experimental phase were a different set of randomly generated sentences (although based on the same exemplars and categories as before).

Each trial consisted of the presentation of a fixation point followed by the presentation of a sentence. Words of the sentence were flashed consecutively onto the screen, one word at a time. Each word remained on the screen for 200 msecs, with a gap of 40 msecs between each word.

The assignment of the right and left hands for True responses was counterbalanced across the experiment. Each experimental session was divided into 3 blocks in order to give participants two moments to pause.

The programs to run all the experiments presented in this paper were written in MATLAB using the Psychophysics Toolbox (Brainard, 1997, and Pelli, 1997).

Results

We analyzed the experiment using both choice proportions and reaction times. In all the analyses using choice proportions, arcsine transformations were carried out before analysis to improve the conformity of the data to the standard assumptions of ANOVA (e.g. Howell, 1997). Likewise, a log transformation was applied to the reaction time data. Following Clark (1973) we also carried out an analysis using both participants and stimuli items as random effects in our model. The item analysis involved summing over all participants but distinguished between the six types of category within our sentences (mammals, birds, insects etc.). We thus had six data points per cell in an analysis of sentence types. By convention, we refer to F-values obtained with participants as the random factor as

participants found T2 Logical sentences more difficult to understand than the other control sentences or the T2 Pragmatic sentences.

F_1 (or t_1), while F-values obtained with items as the random factor as F_2 (or t_2). All p-values assume a two-tailed test unless otherwise stated.

Data treatment. Responses were considered outliers if they were made less than 200 msec after the presentation of the final word or longer than 6 seconds. Outliers were removed from both choice proportions and reaction time data (about 1% of the responses). In addition, when analyzing the reaction time data, we removed all error trials (including those T1 responses that were incorrect with respect to the provided instructions). For example, all trials to which participants gave a True response to Type 6 sentences were removed from analysis. This meant that a further 15% of the responses were removed from the reaction time analysis.

Choice proportion analysis. To compare the proportion of errors made across different stimulus types, we converted proportions of true and false responses to correct and incorrect responses. This means that for T1 sentences, “true” is correct under Logical instructions but “false” under Pragmatic instructions. For the other sentence types, the mapping remains consistent across instructions and is shown in Table 1.

The upper panel of Figure 1 illustrates the proportion correct responses as a function of sentence type and instructions. For the control sentences, approximately 85% of the responses are made correctly and there appears to be little difference between the Logical and Pragmatic instructions. For the underinformative sentences however, there appears to be a large difference between the two conditions: the percentage correct under Pragmatic instructions drops down to 60%, while the percentage under Logical instructions is at the level of the control sentences. This suggests that participants had more difficulty sticking to the instructions in the Pragmatic condition than in the Logical condition.

Insert Figure 1 about here

These observations were verified by running an ANOVA with the transformed proportion correct as the dependent variable, Instructions (Logical or Pragmatic) and Sentence Type (T1-T6) as within-subject factors, and Order of instructions (whether participants were given the Logical or Pragmatic instructions first) as a between subject factor. The interaction between Sentence Type and Instructions was reliable using both participant ($F_1(5,100) = 5.71$; $p < 0.0001$) and item analysis ($F_2(5,25) = 12.28$; $p < 0.0001$). Individual ANOVA's revealed an Instructions by Sentence Type interaction for T1 versus each of the other sentence types; such that T1 sentences were most influenced by the change in instructions (all $F_1(1,20)$'s > 7 , all p 's < 0.02 ; all $F_2(1,5)$'s > 16 ; all p 's < 0.01). There were no effects of Order on responses (all F_1 's < 1.3 , all p 's > 0.28 ; all F_2 's < 1.4 ; all p 's > 0.29) and a disadvantage with respect to correct Pragmatic responses to T1 sentences was present in both orders (Logical then Pragmatic: $t_1(10) = 4.61$, $p = 0.001$; $t_2(5) = 11.57$, $p < 0.0005$; Pragmatic then Logic: $t_2(10) = 4.75$, $p < 0.0001$; $t_2(5) = 6.42$, $p < 0.005$). The results of this analysis indicate that, contrary to a Default Inference account, participants had more difficulty in interpreting *some* to mean *some but not all* than meaning *some and possibly all*.

Reaction time analysis. The lower panel of Figure 1 shows the actual response times to each of the conditions collapsed across order and condition. One can see that the underinformative sentences took longest to process when the instructions encouraged a pragmatic interpretation. The comparisons also show that it is longer than every other condition, including its homologue in the Logical instruction condition. These observations were verified by running an ANOVA with Log(reaction time) as the dependent variable, Instructions and Sentence Type as within-subject factors, and Order of instructions as a between subject factor. The interaction between Sentence Type and Instructions was reliable using both participant ($F_1(5,100) = 7.94$; $p < 0.0001$) and item analysis ($F_2(5,25) = 15.0$; $p <$

0.0001). This confirms that certain sentence types were affected more than others by the instructions manipulation.

To establish whether the T1 sentences were affected most of all by the instructions, we ran individual ANOVA's between T1 and each of the other sentence types (Order was also included as a factor). A reliable difference was observed between Sentence Type and Instructions for each of the comparisons (all $F_1(1,20)$'s > 8.8 , all p 's < 0.001 ; all $F_2(1,5)$'s > 11 , all p 's < 0.05). This demonstrates that even if there is a general effect of the instructions across all sentence types, the inference adds extra processing time. Finally, we examined the extent to which responses to T1 sentences under pragmatic instructions required more time than responses to other sentences under the same instructions. T1 responses were significantly slower than responses to T2, T3, T4 and T6 under pragmatic instructions (all $t_1(21)$'s > 2.2 , all p 's < 0.05 ; $t_2(5)$'s > 2.6 , all p 's < 0.05). The comparison with T5 narrowly failed to reach conventional significance levels using a participant analysis ($t_1(21) = 2.02$, $p = 0.0562$) but was significant with an item analysis ($t_2(5) = 2.8$, $p = 0.038$). To verify that the increase in reaction time was not due in some way to the T1 sentence itself, we compared reaction times of T1 sentences in the Logical instruction condition to the control sentences in the same condition. This analysis demonstrated that response times to T1 sentences in the Logical condition were not significantly different from the response times to the control sentences (using a one tailed test: all $t_1(21)$'s < 1 , all p 's > 0.2 ; all $t_2(5)$'s < 1.4 , all p 's > 0.1).

There was a main effect of Order on the responses ($F_1(1,20) = 6.058$; $p = 0.023$; $F_2(1,5) = 871$; $p = 0.0002$) such that those participants who received the Pragmatic instructions first responded more quickly than those who received the Logical instructions first. The three-way interaction of Sentence by Instructions by Order was also significant, ($F_1(5,100) = 2.5$, $p = 0.035$; $F_2(5,25) = 3.33$; $p = 0.019$). Inspection of the data revealed that response time differences between the two Instruction conditions are largest when

participants see the Pragmatic instructions first. However, this effect failed to reach significance when the Order by Instructions interaction was tested on the T1 items only ($F_1(1,20) = 3.60, p = 0.072; F_2(1,5) = 4.72; p = 0.082$). There was a slowdown for pragmatic responses to T1 sentences in both orders individually, although this effect was not significant in the item analysis of the Logical then Pragmatic instructions order (Logical then Pragmatic: $t_1(10) = 2.57, p = 0.05; t_2(5) = 1.84, p = 0.12$; Pragmatic then Logic: $t_2(10) = 12.63, p < 0.0005; t_2(5) = 8.09, p < 0.0005$).

Discussion

A default view of scalar inference would predict that under Pragmatic instructions, responses to T1 sentences would require less time than responses under Logical instructions. According to an account based on Relevance Theory, one should find the opposite. The data more readily support the Relevance account.

When participants were under instruction to draw the inference, they required more time to evaluate the underinformative sentences than when they were under instructions to provide a logical response. We demonstrated that the underinformative sentence is the one most affected by the instructions. This means that, although the Pragmatic instructions might have increased the difficulty of the task overall, at least some of the extra time required was due to the processing requirements of making and maintaining the inference. This much confirms Rips's initial findings. There are no indications that turning *some* into *some but not all* is effortless.

We also examined whether responses to sentences that required an inference took longer than responses to control sentences under the same instructions. Our results demonstrated that under Pragmatic instructions, this was indeed the case but under Logical instructions no differences were observed. This comparison provides further evidence that the

scalar inference reliably adds processing time that goes above and beyond what is needed for a logical interpretation.

Experiment 2

A potential criticism of Experiment 1 is that the pragmatic effect might be due to a response bias because the correct response to T1 sentences in the Logical instructions condition is to say “True” while under Pragmatic instructions the correct response is to say “False.” If one supposes that people are slower at rejecting a sentence than confirming it, then this alternative explanation predicts an advantage for Logical responses over Pragmatic responses. One response to this criticism is to point out that Pragmatic responses to T1 sentences were not only slower than Logical responses to T1 but also less accurate and slower than the three control sentences that also require a “False” response (T3, T5, & T6); this indicates that T1 Pragmatic responses are exceptional – they are particularly slow and prone to error. Another is to point out how the Logical response to T1 sentences appeared comparable to the control problems (producing rates of correct responses that were indistinguishable from the control problems while being neither exceptionally fast nor exceptionally slow). An even better reply, however, is to allay concerns of a response bias by demonstrating experimentally that the effects linked to pragmatic effort, as exemplified in Experiment 1, are not simply due to hitting the “False” key, the response that was intrinsic to the Pragmatic response in the prior experiment.

Our approach was to work within the same paradigm but to modify it so that the same overt response could be compared across both Logical and Pragmatic instructions; this way, participants’ response choice could not explain the observed effects. In order to arrive at this comparison, we presented two statements per trial: The first one (which we call the “Mary says” sentence and represents the innovation made to the paradigm) makes a True/False declaration about the second (which is one of the 6 types of sentences). The participant’s

task is to agree or disagree with Mary's declaration. By manipulating Mary's declaration about the test sentences, we were able to present trials where the correct response to T1 sentences is "agree" in both the Logical instructions condition and the Pragmatic instructions condition. For example, if a participant in the Logical condition is presented with the sentences, "Mary says the following sentence is true" / "Some elephants are mammals", then the correct response is "agree", because, according to the logical instructions, Mary is correct in saying that the sentence is true. Similarly, if a participant in the Pragmatic condition is presented with the sentences "Mary says the following sentence is false" / "Some elephants are mammals", they should also answer "agree" because, according to the Pragmatic instructions, the sentence is indeed false.

Three variables were therefore manipulated in the experiment. Two of these were present in Experiment 1: the instructions given to participants (either Logical or Pragmatic) and the category sentence type. In addition to these, we manipulated whether the participant should agree or disagree with the "Mary says" declaration. The general expectation is a slowdown whenever the inference is called for and, if the results from Experiment 1 are taken to be indicative, we would expect that the T1 Agree responses in the Logical instructions condition to appear ordinary (because neither the instructions nor what Mary says incite an inference) while the T1 Agree responses in the Pragmatic instructions condition ought to appear exceptionally slow (because the instructions require the production of the inference). Our prediction then is that participants in the Logical instructions condition will correctly respond "agree" more accurately and quickly to T1 sentences than those in the Pragmatic-instruction condition, lending support to a contextualist account of inference generation. We also argue that because participants are making the same overt response across both conditions, a response bias explanation is not a plausible alternative.

We focus exclusively on the “agree” responses because the “disagree” responses, when they are anticipated, arguably require the production of the inference in both the Logical and Pragmatic instruction conditions. One “disagree” response to a T1 statement arises when the Pragmatic-instruction condition necessitates the production of the inference (Mary says “true” and the pragmatic instructions indicate “false”). The other arises in the Logical-instruction condition (Mary says “false” and the instructions indicate “true”). The inference could be prompted here if participants attempt to justify Mary’s declaration. This issue does not crop up in the analysis of “agree” responses and we thus keep our focus on this simpler case.

Method

Participants. Twenty-nine participants were recruited from the Université Catholique de Lyon. All were native French speakers and participated as part of an extracurricular activity for their Introductory Psychology course.

Stimuli and Design. Participants were randomly assigned to one of two groups. In one, participants were told to interpret *some* in a pragmatic way and in the other they were told to treat *some* logically (see Experiment 1). In each group, participants saw the appropriate instructions and went through a practice phase which included feedback. There were 13 participants in the Logical group and 16 in the Pragmatic group. The experiment took place entirely in French.

As in Experiment 1, the category sentences were generated randomly from a base of 6 categories and 9 exemplars from each of these categories. However, we generated two sets of the stimuli to make a total of 108 items. Half of these were prefaced with “Mary says that the following sentence is true” and half prefaced with “Mary says that the following sentence is false”. Any given exemplar (e.g. bee, salmon, dog etc.) was used twice in the experiment, once in the “Mary says...true” trial and once in the “Mary says...false” trial. The

randomization procedure of the program made it highly unlikely (1/36) that a given exemplar would be part of the same type of sentence twice for a given participant.

The task for the participants was to press the key marked “Agree” if they agreed with Mary’s declaration, or to press the key marked “Disagree” if they disagreed with her declaration. In the results section, we will refer to Agree trials and Disagree trials. Agree trials refer to the situations where the “Mary says” declaration is in agreement with the veracity of the category proposition, while Disagree trials refer to the reverse situation. For example, Agree trials for the T2 sentences are trials where Mary’s declaration was “Mary says ... True” because the T2 sentences are true statements. Similarly, the Disagree trials for T2 sentences were those where Mary’s declaration was “Mary says ... False”. Agree trials involved a “Mary says ... true” statement for sentence types T2 and T4, and a “Mary says ... false” statement for types T3, T5 and T6. For those participants in the Logical condition, T1 Agree trials involved a “Mary says ... true” declaration, while for those participants in the Pragmatic condition, the T1 Agree trials involved a “Mary says ... false” declaration.

Participants saw 32 practice statements concerning categories not tested in the experimental session, for example trees and clothes. These sentences were of the form T1, T2, T4, or T5. Participants also saw 5 dummy sentences at the beginning of the experiment to avoid problems associated with starting the experimental phase. All participants saw exactly the same practice and dummy sentences. Participants made their response using the computer keyboard and they were given feedback on all trials, consisting of the word ‘incorrect’ when they responded inappropriately. Of course, the feedback remained the same across conditions for sentences T2-T6. For T1, however, the feedback depended on the type of instructions received.

Procedure. The instructions in the Logical and Pragmatic conditions were the same as those in Experiment 1, with the addition of a few lines explaining the “Mary says”

component. Each trial consisted of the presentation of a fixation point which indicated where the beginning of the test sentence would later appear. This was followed by the “Mary says” sentence (“Mary says that the following sentence is true/false”), which remained on the screen for two seconds before the test sentence appeared and roughly 2 centimeters above the eventual test sentence. The test sentences in this experiment were presented in full (i.e. not one word at a time). Both the “Mary says” sentence and the test sentence remained on the screen until the participant responded. The assignment of the right and left hands for Agree responses was counterbalanced across the experiment. Each experimental session was divided into 3 blocks in order to give participants two breaks.

The procedure used for practice trials was identical to the experimental trials. However, participants were encouraged to ask questions during the practice phase and to work independently during the experimental session. Participants were not told of the existence of the dummy sentences.

Results

As in Experiment 1, we analyzed the results using both choice proportions and reaction time, with participants and items as random factors. We concentrate primarily on the trials where an Agree response was required by the participants, that is, trials where the “Mary says” declaration is in agreement with the category proposition. This was because we believed participants might make the inference in both the Logical and the Pragmatic instructions conditions on Disagree trials. Nonetheless, we present a brief summary of the Disagree responses at the end of the Results section.

Data treatment.

Responses with an associated reaction time of less than 0.5 seconds or more than 25 seconds were considered outliers and removed from further analysis. This eliminated less than 0.5% of the responses. Outlier limits were different to the previous experiment because

this task was considerably more difficult and reaction times were consequently much higher. When we performed the analysis on reaction times, we also removed incorrect responses as we did for the previous experiment. This resulted in a further 11% of the data being eliminated.

Choice proportion analysis of Agree responses.

The upper panel of Figure 2 shows the proportion of correct Agree responses as a function of the type of instructions presented. The general pattern of responses validates the findings from Experiment 1, with low accuracy for the Pragmatic T1 sentences compared to both Logical T1 sentences and the other control sentences. As before, this pattern of results does not support the hypothesis that the *some but not all* interpretation is the default or automatic interpretation. With the transformed proportion correct as the dependent variable, an ANOVA was conducted with Instructions (Logical or Pragmatic) and Sentence Type (T1-T6) as factors. There was no main effect of the Instructions ($F_1(1,27) = 0.071, p = .791$; $F_2(1,25) = 0.079, p = .789$), but the interaction between Instructions and Sentence Type was reliable ($F_1(5,135) = 3.571, p = .005$; $F_2(5,25) = 4.042, p = .008$). To establish where these effects were, we conducted t-tests on each of the sentence types. On T1 sentences, responses under Pragmatic instructions were reliably less accurate than responses under Logical instructions ($t_1(27) = 2.522, p = .018$; $t_2(5) = 2.8; p = .038$), suggesting that the Pragmatic interpretation is more difficult than the Logical interpretation, even when the overt response made by participants was identical. Furthermore, there were no reliable effects of the instructions manipulation among the control sentences, thus eliminating the possibility that there was a general effect due to instructions (t_1 's(27) < 2.04, p 's > .05; t_2 's (5) < 2.5; p 's > .05; except for T5 in the participants analysis: $t_1(27) = 2.3, p = .028$, but after adjustment for multiple comparisons this effect is no longer reliable).

We also compared T1 sentences to the control sentences within each of the instructions conditions. This was to establish whether there was something about the structure of T1 sentences, apart from the inference, that made them particularly difficult to interpret. Comparison of T1 sentences against T2-T6 sentences in the Logical instruction condition revealed no reliable differences (all $t_1(12)$'s < 1 , p 's $> .4$; $t_2(5)$'s < 0.9 , p 's > 0.4), whereas comparisons within the Pragmatic condition revealed that responses to T1 were significantly less accurate than those of the control sentences ($t_1(15)$'s > 2.15 , p 's $< .05$; $t_2(5)$'s > 3.01 , p 's < 0.05 ; except for the comparison with T2, $t_1(15) = 1.98$, $p = .066$; $t_2(5) = 2.2$, $p = .079$). Thus, as in Experiment 1, we can conclude that there is nothing unusual about T1 sentences themselves, but that it is the addition of the inference that causes the drop in accuracy.

One confound associated with analyzing the T1 Agree responses is that the “Mary says” declaration is “Mary says false” in the Pragmatic instructions condition while it is “Mary says true” in the Logical condition. The presence of the word “false” in the Pragmatic condition might therefore be responsible for the drop in accuracy. However, if this were the case, then we would expect no differences between responses to T1 Agree sentences in the Pragmatic condition and the control sentences involving the “Mary says false” declaration, i.e. T1 responses should not be different from T3, T5 and T6. Furthermore, we would expect the T1 Logical responses to be more accurate than the T3, T5 and T6 Logic controls because T1 Logic responses involve a “Mary says ... true” declaration whereas these control sentences involve a “Mary says ... false” declaration. As we demonstrated in the paragraph above, neither of these predictions were borne out so we can reject the hypothesis that the low accuracy for T1 Pragmatic sentences was due to the “Mary says ... false” declaration.

A similar criticism is that the transition between evaluating a true proposition and providing an “agree” response might be easier than making the transition between evaluating

a false proposition and providing an “agree” response, hence the difference between the Pragmatic and Logical instructions conditions on T1 sentences. To test this hypothesis, we compared Agree responses to the control sentences that involved true propositions (T2 and T4), with Agree responses that involved false propositions (T3, T5, and T6). If the ease of transition between veracity of proposition and response type was responsible for the difference on T1 sentences, then we would expect to find similar effects on the control sentences. In the event, no reliable differences were observed between the two sets of control sentences, $F_1(1,29) = 1.086, p = .307$; $F_2(1,10) = 0.91, p = .364$.

 Insert Figure 2 about here

Reaction time analysis of Agree responses.

The lower panel of Figure 2 displays the time taken to correctly respond Agree as a function of the type of instructions received. As with the choice proportion analysis, the results replicate those of Experiment 1; T1 Pragmatic responses appear to require more time than T1 Logical responses or the control sentences, even though all participants are now responding with Agree responses.

We analyzed the results using an ANOVA with the log-transformed Reaction Time to the correct Agree response as the dependent variable, and Instructions and Sentence Type as the two factors. There was a main effect of the Instructions using items as a random factor ($F_2(1,25) = 7.295, p = .043$), but not participants ($F_1(5,135) = 0.234, p = .63$). However, the interaction between Instructions and Sentence Type was significant using both items and participants ($F_1(5,135) = 6.419, p < .0001$; $F_2(5,25) = 3.245, p = .022$). A t-test between T1 Logical and T1 Pragmatic revealed that T1 Pragmatic responses were reliably slower, ($t_1(27) = 2.82, p = 0.009$; $t_2(5) = 4.076, p = .01$). Furthermore, the effect of the instructions was

limited to T1 sentences, as indicated by the comparisons with the control sentences (all $t_1(27)$'s < 0.7 , p 's > 0.4 ; all $t_2(5)$'s < 1.3 , p 's > 0.25).

We were also interested in identifying whether the T1 sentences were generally difficult to process. To this end, we compared the T1 responses to the other sentences within the two Instructions conditions. Among the Logical instruction responses, there were no reliable differences between T1 and the control sentences ($t_1(12)$'s < 1.47 , p 's > 0.15 , all $t_2(5)$'s < 2.24 , p 's < 0.05 ; except for T1 vs T5: $t_1(12) = 2.269$, $p = .043$, where adjustments for multiple comparisons make the comparison unreliable). In contrast, comparisons among Pragmatic responses revealed that T1 responses were reliably slower than all of the control sentences ($t_1(15)$'s > 2.733 , p 's $< .02$; $t_2(5)$'s > 2.55 , p 's $\leq .05$). These results demonstrate that there is nothing about the T1 sentences themselves that are difficult to process; rather, the introduction of the scalar inference causes the slowdown in its interpretation.

In our analysis of the choice proportions, we considered the possibility that providing an “agree” response to a statement involving a true proposition might be easier than providing an “agree” response to a statement involving a false proposition, and that this could then explain the differences observed on our T1 sentences. A similar proposal could be made to account for the reaction time data. As before, we tested this by comparing control sentences that involved a true proposition with control sentences that involved a false proposition. Such an analysis performed using reaction times again failed to produce any reliable effects, $F_1(1,27) = 0.323$; $p = .574$, $F_2(1,10) = 0.02$, $p = .89$, suggesting that the ease of response transition is unlikely to account for slower reaction times to the T1 Pragmatic sentences.

Disagree responses.

There appeared to be no differences between the Pragmatic and Logical conditions within the disagree responses. ANOVAs were conducted with Instructions and Sentence

Type as factors, and either choice proportions or reaction time as dependent measures. No reliable effects involving the Instructions factor were present ($F_1(5,135)$'s < 1 , p 's $> .5$; $F_2(5,25) < 2$, p 's $> .1$), although there were main effects of Sentence Type using both choice proportions and reaction time as dependent measures (F_1 's > 10 , p 's $< .0005$; $F_2(5,25)$'s > 3.8 , p 's < 0.01). As an extra check, we performed t-tests on the T1 sentences between the Logical and Pragmatic instructions conditions. No effects were observed when choice proportions were used with participants as the random factor ($t_1(27) = 0.326$, $p = .75$), although when items was used the comparison approached significance ($t_2(5) = 2.5$, $p = .057$) such that those in the Pragmatic condition were less accurate than those in the Logical condition. Similarly, no effects were observed when reaction times were used as the dependent measure ($t_1(27) = 0.316$, $p = 0.76$; $t_2(5) = 0.665$, $p = .54$).

In summary, the analysis of the disagree responses supports our initial predictions that participants would make the inference in both Logical and Pragmatic conditions for the T1 trials. Furthermore, the failure to find any effects of the Instructions factor supports our claim that the effects observed in the Agree responses were due to the inference and not some general instructions effect.

Discussion

This experiment verified that inaccurate and slow T1 Pragmatic responses are not due to the provided response options. By manipulating the “Mary says” declaration preceding the category sentences in this experiment, we were able to compare participants making the same overt response to T1 sentences in the Logical and the Pragmatic conditions. Our results demonstrate that participants were slower and less accurate in correctly agreeing with T1 when given instructions to treat *some* pragmatically rather than logically.

Correct response patterns validate those seen in the previous experiment, where responses indicating a Logical treatment of T1 were made accurately and quickly, much like

the control items, while responses indicating a Pragmatic treatment of T1 were more error-prone and exceptionally slow. The relatively low rates of accuracy, and concomitant slow speeds, linked with agreeing with T1 sentences in the Pragmatic instruction condition are exceptional, not only when compared to T1 in the Logical condition but, when compared to the control items. The analysis of correct “agree” responses confirms the conclusion from Experiment 1 - responses that integrate a scalar inference require exceptional effort to be processed. These findings demonstrate that the pragmatic effect reported in Experiment 1 cannot be explained by a response bias.

We now turn to another potential criticism of our first two experiments, which is that by giving participants explicit instructions about the interpretation of *some*, we might be asking them to use the word in a way that goes counter to their own predilections. Perhaps participants see the quantifier and ask themselves which is the appropriate meaning of the word, rather than directly process its meaning. Although we do not see why this would lead to differences between the Logical and Pragmatic conditions (rather than just adding noise in general), we feel it is appropriate to run another experiment that is more ecologically valid.

Experiment 3

This experiment uses the same paradigm as in Experiment 1, however we provide neither explicit instructions nor feedback about the way to respond to T1 sentences. Instead, we expect participants’ responses to reflect equivocality to these types of sentences - some saying false and some true. This means that we should have two groups of responses: one in which the inference is drawn (T1 Pragmatic responses) and another where there is no evidence of inference (T1 Logical responses). We can therefore make a comparison between the two as we did in the prior experiments. Once again, if logical responses are made more quickly than pragmatic responses, we have evidence against a default system of inference. We can also use the control sentences to verify that under such neutral instructions, responses

which involve the inference require more time than responses that do not (as we found in Experiment 1).

Method

Participants. Thirty-two undergraduates from the Université de Lyon 2, who were either volunteers or presented with a small gift worth about 5 Euros, participated in this study.

Stimuli and Design. There was no instructional manipulation in this experiment so participants went through only one experimental session. As before, participants saw 9 examples of 6 types of sentences, making a total of 54 experimental items. The stimuli were generated in the same way as for Experiment 1. No practice session was given because participants no longer had to automate specific instructions, although dummy sentences were still presented at the beginning of the experiment.

Procedure. Participants were placed in front of a computer and told that they would see sentences presented on the screen. In contrast to the previous experiment, the only instructions they were given was to respond 'True' if they thought the sentence on the screen was true, or 'False' if they believed the sentence to be false. Participants were not told whether their responses were correct or incorrect, i.e. there was no feedback.

Each sentence was presented in its entirety on the screen. The sentence remained on the screen until the participant made a response. All other aspects of the experiment were identical to Experiment 1.

Results

Data treatment. Outliers were considered to be responses made in less than 0.5 seconds or more than 10 seconds. This resulted in less than 1% of the trials being removed from the data set. Note that the criteria for removing data points appear different from those of Experiment 1; this is because the participants here had to read an entire sentence within

this time period (in Experiment 1, participants were timed from the moment the last word of the item appeared). For the purposes of our analyses of reaction times, we include only correct responses (among the control sentences, Type T2-T6). This resulted in an additional 10% of the responses being removed. For Type T1, both types of responses are justifiable and are included.

Analysis of choice proportions. The nine individual trials for each sentence type were pooled, producing a set of six means per participant. Means and variance of the response types are shown in Table 2 as a function of sentence classification and stimulus type. In the 5 control sentences, participants were largely in agreement in choosing true or false responses. Correct responses for T2 through T6 ranged from 87% to 98%. As demonstrated elsewhere (Noveck 2001), responses to underinformative sentences prompt a high degree of bivocality - 61% of responses here were pragmatic interpretations. The difference in variability between T1 sentences and each of the control sentences was confirmed by performing Levine's test of equal variances (on the untransformed proportions): the variance of the T1 sentences is significantly higher than any of the other sentences types, with all p 's < 0.0001.

 Insert Table 2 about here

Analysis of reaction times. In order to assess whether a logical response was made more quickly than a pragmatic response, we divided each participant's answers to T1 sentences into Logical and Pragmatic and then found the mean reaction time for these two groups (see Figure 3). This gave us a within-participant measure of the change in reaction time for response type. Nine participants were excluded from the main analysis because they responded with a single type of response – either all Logical (2) or all Pragmatic (7) – and were thus ineligible for a repeated measures analysis. We discuss these nine participants at

the end of this section. A paired t-test revealed that the time taken to respond Pragmatically to T1 sentences takes significantly longer than the time taken to respond logically to T1 sentences ($t_1(22)=2.07$, $p = 0.05$; $t_2(5) = 4.7$, $p = 0.0054$).

Insert Figure 3 about here

We also carried out tests that compared the control items to each of the two kinds of responses to T1 in order to determine whether the Pragmatic responses are characteristically different than the other responses in this task. If scalar inference-making is unique to responses to T1, it implies that such responses should take longer than responses to each of the control sentences (again, note that three of these - T3, T5, and T6 - also require a False response). Paired t-tests between T1 Pragmatic responses and control sentences reveal that this is indeed the case for nearly all of the control sentences ($t_1(22)$'s > 2.1 , p 's < 0.05 ; $t_2(5)$'s > 3.2 , p 's < 0.05). The only difference which failed to reach significance levels was that between T1-Pragmatic and T4 using items as a random factor ($t_1(22) = 2.11$, $p < 0.05$; $t_2(5) = 2.443$, $p = 0.058$). However, because this comparison was found to be significant in the previous experiments, we have confidence in its reliability. (Note that the difference between T1 and T5 which was narrowly non-significant in Experiment 1 has been demonstrated reliable in this experiment.)

To further ensure that longer reaction times to T1-Pragmatic responses were not just due to the difficulty in interpreting the sentence structure of T1, we compared the Logical responses to T1 to the responses to each of the control sentences. If the T1 sentences were in some way characteristically different from the other sentences, one would expect that even those who gave Logical responses to T1 sentences to have longer reaction times than they did to control sentences (which include two, T2 and T4, requiring a True response). This is not

the case. Although the comparison between T1 Logical and T6 was found to be reliable using participants as a random factor (using a one-tailed test: $t_1(22) = 2.37$, $p = 0.014$; $t_2(5) = 1.23$, $p = 0.14$), practically all comparisons showed no reliable differences (all $t_1(22)$'s < 1.1 , all p 's > 0.13 ; all $t_2(5)$'s < 1 ; all p_2 's > 0.25). Thus, we feel safe in concluding that at least some of the extra time required to respond pragmatically to T1 sentences is linked to the added effort of making the inference.

As indicated above, nine participants were removed because they responded with a single type of response and were thus ineligible for a repeated measures analysis. These participants were similar to those who gave two kinds of responses over the course of nine trials; logical participants responded more quickly to T1 sentences than pragmatic participants. To verify this, we ranked the participants in terms of mean reaction times. The two participants who responded logically have the two lowest reaction times out of the 9. This leads to a two-tailed p -value of 0.1 for a Wilcoxon Signed-Ranks test (the lowest possible for this ratio of participant numbers, $W_s = 3$, $n_1 = 2$, $n_2 = 7$). In sum, there is no evidence whatever that the removal of these participants biased the analysis of the experiment.

Discussion

The main finding here is that mean reaction times were longer when participants responded pragmatically to the underinformative T1 sentences than when they responded logically. Furthermore, pragmatic responses to the underinformative sentences appear to be slower than responses to all of the control sentences, indicating that the scalar inference, which is unique to Pragmatic responses to T1, prompts an evaluation that is characteristically different from all the other items. These results are shown to occur with both a participant analysis and an item analysis. Collectively, our findings provide further evidence against the default inference view because there is no indication that participants require more time to

arrive at a true response for the T1 sentences than they do to a false response. All indications point to the opposite being true: A logical response is an initial reaction to T1 sentences and it is indistinguishable from responses to control sentences while a pragmatic response to T1 is significantly slower than a logical response to T1 and to the other items in the task.

As we argued in the prior experiments, the exceptional nature of the T1-Pragmatic response cannot be attributed to the false response it engenders because the response to this sentence is also slower than all three of the control sentences that require a false response. Consider T5 which also mentions a category and its member (e.g. *All mammals are elephants*) and *requires* a False response. Such items prompt 97% of participants to respond False correctly and at a speed that is significantly faster than it is for the T1-Pragmatic responses. Similarly, it cannot be argued that false responses to T1 sentences are due to error (meaning that participants intended, but failed, to hit the True key) because the percentage of participants making T1-Pragmatic responses is of a characteristically different order when compared to those in the control conditions (roughly 60% choose False to T1 sentences as opposed to 3-13% who make errors across all the control conditions). We argue that these results indicate that the scalar inference is at the root of the extraordinary slowdown in this paradigm. It is drawn specifically in reaction to the underinformative (T1) items and prompts participants to ultimately choose False. Furthermore, it arrives as a secondary process relative to a justifiable logical interpretation; it does not appear to arrive by default.

Although our experiments provide evidence against the idea that scalar inferences become available as part of a default interpretation, they do not necessarily provide evidence in direct support of the alternative presented here, the Relevance theory explanation. Our goal in the next experiment is to test directly predictions from Relevance theory concerning the processing of scalar inference.

Experiment 4

According to Relevance theory, inferences are neither automatic nor arrive by default. Rather, they are cognitive effects that are determined by the situation and, if they do manifest themselves, ought to appear costly compared to the very same sentences that do not prompt the inference. In Relevance terminology, all other things being equal, the manifestation of an effect (i.e. the inference) ought to vary as a function of the cognitive effort required. If an addressee (in this case, a participant) has many resources available, the effect ought to be more likely to occur. However, if cognitive resources are rendered limited, one ought to expect fewer inferences. Experiment 4 tests this prediction directly by varying the cognitive resources made available to participants. The experiment follows the general procedure of Experiments 1 and 3, in that participants are asked to judge the veracity of categorical statements. The crucial manipulation is that the time available for the response is varied; in one condition participants have a relatively long time to respond (referred to as the Long condition), while in the other they have a relatively short time to respond (the Short condition). By requiring participants to respond quickly in one condition, we intend to limit the cognitive resources they have at their disposal. Note that it is only the time to *respond* which is manipulated; participants are presented with the words one word at a time and at the same rate in both conditions, thus there is no possibility that participants in the Short condition spend less time reading the sentences than those in the Long condition.

We wished to make the Long condition as much like the previous experiments as possible. This meant that we chose a response lag duration which we believed would not put participants under any pressure to respond quickly but nonetheless kept the idea that they had to respond within a certain time limit. Judging from previous experiments, three seconds appeared to be ample time to make the response. In contrast, we wanted participants in the Short condition to have sufficient time to respond but to feel under time pressure. We

therefore set the duration of the short lag to be approximately equal to the mean reaction time across the Logical condition of Experiment 1 (900 msec). We felt that a lag time shorter than this would result in too many error responses while a longer lag would not exert enough time pressure.

Relevance Theory would predict fewer inferences when participants' resources are limited. It is expected that they would be more likely to respond with a quick "True" response when they have less time than when they have more. If one wanted to make predictions based on the DI approach, *some* should be interpreted to mean *some but not all* more often in the short condition than in the long condition (or at least there should be no difference between the two conditions).

Method

Participants. Forty-five participants from the Université de Lyon 2 were used in the study. Participants were either volunteers or were presented with a small gift worth about 5 Euros.

Stimuli and design. Participants again had to respond true or false to 54 category statements, generated in the same way as in Experiment 3. Participants were given the same 16 practice sentences as described in Experiment 1, as well as the dummy sentences before the experiment. The new independent variable was the time that participants were given to respond to the statement, referred to as the Lag. The Lag was a between participant variable which could be either a short time (900 ms) after the presentation of the final word, or a long time (3000 ms). The dependent measure was the proportion of true responses within the time lag. Twenty participants were assigned to the short lag and 25 to the long lag.

Procedure. The instructions for both conditions were similar to those of the previous experiment. Participants were told that they would see sentences presented one word at a time on the screen and that they would have to say whether they considered the sentences to be

true or false. They were not given specific instructions on how to interpret *some*. In both Long and Short conditions, participants were instructed that if they took too long to respond they would see a message informing them of this. In the Short condition, speed of response was emphasized and participants were told that they would have to respond in less than half a second. We chose to lower estimates to half-a-second for the instructions because hitting the response key was expected to take up a portion of the 900 msec. In any case, training gave participants a clear idea of how much lag time is available in the Short condition.

Sentences were presented one word at a time on the screen, in the same manner and for the same length of time as in Experiment 1. We chose this method (instead of presenting the whole sentence at once) because we wanted to make sure that participants in both conditions spent an equal amount of time reading the words. This was to stop people in the Short lag condition from simply scanning the sentence and basing responses on the most salient components in the sentence.

The trial by trial procedure was identical to that of Experiment 1 until the participant made their response. After the response, the participant was told whether they were 'in time' or 'too slow'. In the Short condition they were 'in time' if they responded in less than 900 ms, whereas in the Long condition the limit was 3000 ms. The timing feedback remained on the screen for 1 second. Participants were not given feedback on the whether their response was correct or not.

Results

Data treatment. Responses that were outside the allotted time lag for each condition were removed from the analysis. Thus, responses were removed if they had an associated reaction time of more than 900 ms in the Short condition and more than 3000 ms in the long condition. This resulted in a total of 12 % eliminated from the Short condition and 0.7% from

Long condition. There appeared to be a uniform distribution of removed responses across the different sentence types.

Analysis. Table 3 shows the rates of True responses for all six sentence types. The rate of correct performance among the control sentences either improves (T3 - T6) or remains constant (T2) with added response time. This trend is shown in the last column of Table 3 which, for control sentences, indicates the increase in proportion correct with added response time. In contrast, responses to the underinformative sentences were less consistent with added time available. This change was such that there were more Logical responses in the Short condition than in the Long condition: 72% True in the Short condition and 56% True in the Long condition. This trend is in line with predictions made by Relevance theory.

Insert Table 3 about here

To confirm these observations, we ran an ANOVA with Sentence Type and Lag as factors and proportion of True responses as our dependent measure. This revealed a significant interaction of Lag by Sentence Type ($F_1(5,215) = 2.549$, $p = 0.039$; $F_2(5,25) = 2.63$, $p = 0.049$). To discover which sentences were affected by the lag factor, we ran individual t-tests between the two lag conditions. We had *a priori* predictions that there would be more Logical responses in the Short lag condition for T1 sentences but no other predictions. T1 sentences showed a reliable difference between the two lags ($t_1(43) = 2.43$, $p = 0.019$; $t_2(5) = 6.6$, $p < 0.001$ assuming one-tailed tests), and there was some evidence of differences on T4 sentences ($t_1(43) = 2.21$, $p = 0.032$; $t_2(5) = 1.17$, $p = 0.30$ assuming two-tailed tests), although after correcting for multiple comparisons the results do not come out significant. No other sentence types differed (all p 's > 0.1 ; all p_2 's > 0.07).

Below, we compare performance on the T1 sentences in the Short condition to two sorts of chance conditions, one in which chance is 0.5 and another which is based on a stricter determination of chance conditions. To test the first, we performed a one sample t-test in order to determine whether the percentage of Logical responses to T1 sentences in the Short condition was significantly greater than a traditional interpretation of chance (0.5), $t(19) = 4.3$, $p < 0.001$. This indicates that participants were unlikely to have responded “True” by chance alone.

Although participants were not responding *entirely* by chance in the Short condition, it is possible that some participants made errors when they could have benefited from more time. In other words, it is conceivable that some proportion of the participants – i.e., those who would have taken longer than 900 msec to answer under more ideal conditions and who would have been ultimately “pragmatic” -- were destabilized by the short lag and responded randomly. This could explain the pattern of results in the Short condition without being due to a deliberate response pattern. We thus calculated an adjusted chance level, which was determined as follows. First, we looked at the distribution of reaction times in the Long condition and found that 43% of the responses were below 900 msec (the duration of the short lag). Of these, 74% were Logical and 26% Pragmatic (i.e., of the 43% under 900 msec, 32% of the total were Logical and 11% Pragmatic). This means that we would expect at least 32% of the responses under the Short condition to be Logical and 11% Pragmatic because these would not be affected by the short time lag. Under this adjusted-chance procedure, the remaining 57% would be made by chance and would therefore consist of 28.5% True and 28.5% False responses. If we add the 28.5% True to the 32% Logical (and we round up), we arrive at a figure of 61%. This represents a more severe estimate of the mean percentage of true responses with an adjusted chance level. As before, we then carried out a one sample t-test against the null hypothesis that our sample came from a population

with a mean of 61%. The t-test confirmed that the observed rate of “true” responding (72%) was significantly different from 61%: $t(19) = 2.6, p < 0.02$. We can thus reject the notion that the 72% figure is the result of some combination of chance responses that arise due to those participants who are being blocked from giving a Pragmatic response. We can thus conclude with greater confidence that Logical responses are being made deliberately as a result of the limited time, as would be predicted by Relevance theory.

Discussion

This experiment manipulated the time available to participants as they were making categorization judgments. We found that when a short period of time was available for participants to respond, they were more likely to respond “True” to T1 sentences. This strongly implies that they were less likely to derive the inference when they were under time pressure than when they were relatively pressure-free. Furthermore, we eliminated the possibility that the difference between conditions can be in any way due to chance responding.

The control sentences provide a context in which to appreciate the differences found among the T1 statements. They showed that performance in the Short Lag condition was quite good overall. In fact, the 72% who responded “True” in T1 represented the lowest rate of consistent responses in the Short condition. All of the control sentences in both the Short and Long lag conditions were answered correctly at rates that were above chance levels. For the control sentences, correct performance increased with added time. This experiment confirms a very specific prediction of Relevance Theory - a reduction in the cognitive resources available reduces the likelihood that the scalar inference will be made.

General Discussion

The experiments presented in this paper were designed to compare two competing accounts about how scalar inferences are generated. Participants were asked to evaluate

statements that could be interpreted in one of two ways: either by treating the quantifier *some* in a logical way and not attaching any inference or by drawing a scalar inference and treating *some* to mean *some but not all*. The theories under consideration make different predictions regarding the length of time required to make the different responses. A Default Inference account would predict that a logical interpretation would take longer than a pragmatic interpretation because the inference would first have to be cancelled before the weaker sense of the word was processed. Relevance Theory would argue that inferences arise as a function of effort; weaker interpretations (in this case, logical ones) could serve initially for providing a response. Thus, according to Relevance Theory, the logical response ought to be faster than a pragmatic response.

In Experiment 1, we gave explicit instruction about the way the weak quantifier *some* ought to be interpreted. A within-participant study showed that those participants who were given instructions to treat *some* as *some and possibly all* responded more quickly to underinformative sentences than those who were given the instructions to treat *some* as *some but not all*. When participants said “true” in the Logical instruction case, their responses and their speed in responding were indistinguishable from the control sentences. Moreover, error rates in the Logical instruction condition were significantly lower among the underinformative sentences than in the Pragmatic instructions condition, indicating greater ease in treating the underinformative sentences in a logical guise. In contrast, when the same participants were asked to treat *some* as *some but not all*, reaction times slowed down significantly for the underinformative sentences. These findings, which largely confirm a result from a very early study by Rips, lend doubt to a Default Inference account that the initial treatment of *some* is *some but not all*. In Experiment 2, we altered the design of the experiment so that identical overt responses would be made across both the Logical and the Pragmatic conditions. Our findings supported the results obtained in Experiment 1, thus

eliminating doubts that our results could have been due to a response bias that favors “true” and disfavors “false”. In Experiment 3, there were no specific instructions about the meaning of *some* as participants were free to respond “true” or “false” to the provided statements. Responses from participants in this investigation again indicated that a pragmatic interpretation to the underinformative sentences were exceptional slow, taking longer than the logical interpretation and the control sentences. As in Experiment 1, there is no indication that *some but not all* is the interpretation of *some* by default.

Experiment 4 presented a more direct test of the Relevance account. Cognitive resources were manipulated (by way of time available for responding) to see whether fewer resources were linked with fewer inferences. In the experiment, those who had less time to respond to underinformative items (900 msec), responded using a logical interpretation at rates that were above chance levels. Meanwhile, they also answered the control items correctly at rates that were even higher. As this account would predict, when resources were made more available by way of increased time (3 seconds), it coincides with more scalar inference production and, thus, higher rates of pragmatic interpretations. All told, the results from the four experiments indicate that people initially employ the weak, linguistically encoded meaning of *some* before employing the scalar inference.

Until now, we have concentrated on theoretical linguistic-pragmatic accounts for the way scalar inferences are drawn out of *some*. Here we consider a psychological possibility, which is that the error rates and slowdowns related to pragmatic readings of *some* results from the nature of the *some but not all* proposition itself. This explanation places the weight of the slowdown not on drawing the inference *per se*, but on the work required to determine the veracity of a proposition with the inference embedded within it. There are two ways in which the *some but not all* proposition is more complex than, say, *some and possibly all*. One is that such a proposition gives rise to a narrower set of true circumstances; thus determining

whether or not a statement is true requires more careful assessments. The other is that negation, as is often the case, adds costs to processing (Just & Carpenter, 1971; Clark & Chase, 1972; although see Lea & Mulligan, 2002). This resonates with the intuition from Rips (1975, p. 335) described in the Introduction, who suggested that the negation in the pragmatic reading of *some* is the source of the slowdown. Both of these suggestions are worthwhile descriptions of the cause of inference-related slowdowns and worth further study. However, neither of these is inconsistent with Relevance theory's account, which makes the original counterintuitive prediction that the pragmatically enriched interpretation requires effort.

One psychological model that could accommodate our findings is Sanford and colleagues' account of non-standard quantifiers (e.g. Sanford, Moxey and Paterson, 1996). As described in the Introduction, their account is not only compatible with the approach we defend here but is enriched by it. The two are compatible because Relevance Theory and the Focus account from Sanford et al. would agree that a) quantifier interpretation relies on attributions of a speaker's expectations and that b) quantifier interpretation is context dependent. Our study with *some* adds an intriguing layer to the Focus account because we investigated a standard positive quantifier that ought to put the focus on the quantified object. This much appears to be the case for those who respond logically. For those who go further, however, a scalar places a focus on the Complement Set, essentially transforming the positive *some* into a negative quantifier. We suggest that such a participant – looking for a more relevant reading – prompts the scalar and notices the null Complement set. For example, a scalar prompted by an underinformative item like *Some monkeys are mammals* puts a focus on the Complement Set (non-mammalian monkeys); when the participant realizes that there is no such thing as non-mammalian monkeys, they respond false. It could very well be that

the search for the non-existent Complement leads to the extraordinary slowdown we report here.

In the Introduction we drew a distinction between the predictions we generated from a Default Inference processing model and the Neo-Gricean theory, as exemplified by Levinson (1983, 1987, 2000). Here we discuss whether it is possible to reconcile a Neo-Gricean view of scalar inference with the results of our experiments. One possibility is to assume that the pragmatic interpretation of *some*, rather than being produced by default and then cancelled, is in some cases preempted. In other words, the theory still incorporates a default inference, but, in some special contexts, the inference is cancelled *before* the scalar term could provoke it. A Neo-Gricean account could then claim that our experiments invoked just such a special context and that the results do not provide evidence against default inferences in the normal situation. Our response to such a point is to first argue that such a contextually-sensitive default theory seriously compromises the usefulness of the default notion in general. The advantage of defaults for the efficiency of processing lies in the automaticity of the default inference; it would be problematic if defaults fail to occur in unforeseen contexts such as the one in our task. Furthermore, if there are numerable cases in which the default evaporates, we argue that the Neo-Gricean account (whether it be defended by Levinson, Chierchia or others) would have to be much clearer about when the default does not apply and it would have to anticipate our results. Our crucial test sentences are unembedded (e.g. they are not in downward entailing contexts nor preceded by clauses like “For all I know”) and are in principle not exceptional according to a neo-Gricean account like Levinson’s or a semantic account like Chierchia’s.

Second, the default mechanism - as it applies to the underinformative statements tested here - does not appear to be categorical in nature. A pre-emption, if it were to occur in a systematic way, ought to apply to all of our underinformative statements (or if there were

no claim for pre-emption, to none of them). Instead, the default mechanism appears to operate in roughly half the cases and in no predictable manner. This lack of systematicity is problematic for a general default mechanism account.

One might be tempted to reconcile our findings with a default account by arguing that the nature of T1 sentences is such that it pre-empts the production of the scalar, leading to a *facilitation* of the Logical interpretation. However, if that were the case, there ought to be evidence indicating that the logical response to T1 is significantly faster than, not only the pragmatic response but, the controls (or at least control sentences T2 and T3, which employ *Some*) and there is nothing in the data to support this prediction. The production of the scalar inference is linked with an extraordinary slowdown among the underinformative items only and it is also slower than the speed of response to the control items; meanwhile, the speed of providing a logical response to T1 items is comparable to the response times of the control items.

Another query concerns our materials: Are the test sentences in our experiments representative of everyday conversation? In our experiments, participants have the choice between two interpretations, neither of which appears compelling or favored, whereas in (2) above, when Robyn replies “Some of them”, it is obvious that the hearer should draw the scalar inference to understand *some but not all*. The point behind this query is that perhaps non-standard sentences imply non-standard conversational strategies.

Our response to this is threefold. First, we point out that the kind of interpretive equivocality we find in our experimental material is not without counterpart in ordinary conversation. Imagine for instance that Henry has, in front of his colleagues, drunk all six bottles of a six-pack. He now concedes: “OK, I have drunk some of them.” Is this to be interpreted as an underinformative statement or as implying that he has not drunk all of the bottles, and therefore a blatant lie? Neither interpretation is compelling or satisfactory, but

either can be accessed by ordinary comprehension mechanisms. Our work has shown that the logical interpretation is fast and that the pragmatic one is exceptionally slow.

Secondly, the query suggests that mechanisms involved in comprehending our crucial test sentences are not the same ones as those used in everyday comprehension of conversation, written texts, exam questions, and so forth. This would be a novel suggestion. To the best of our knowledge, nobody has ever suggested that the cognitive mechanisms handling statements in an experimental situation are different from those applied to ordinary statements in actual verbal exchanges.

Finally, experimental material that elicit two types of interpretations with comparable frequencies, far from being flawed, is in fact optimal for our purpose since it eliminates all factors other than the choice of interpretation as a plausible cause of the time taken. Let us add that the materials in the present study are the result of an evolution in our experimental paradigm that originally used conversational contexts (note the Experimenter-handled puppets and the double blind oral tests in Noveck, 2001). In a nutshell, our experiments involve artificial stimuli for the same reasons most experiments do: these stimuli allow for fine-grained controlled comparisons not available with more real-life material and situations. The pragmatic phenomena we are discussing have been studied mostly on the basis of linguistic intuitions and anecdotal observations. We feel that experimental material of the type we use here, that is, utterances the interpretation of which can go in two different directions, provides crucial evidence for evaluating pragmatic claims.

Conclusion

This work largely validates distinctions made by Grice nearly a half-century ago by showing that a term like *some* has a logical reading and a pragmatic one. This study focused on the pragmatic reading that results from a scalar inference. It does not appear to be general

and automatic. Rather, as outlined by Relevance Theory, such an inference occurs in particular situations as an addressee makes an effort to render an utterance more informative.

References

- Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*(10), 433-436.
- Carston, R. (2002). *Thoughts and Utterances*. Oxford: Blackwell.
- Chierchia, G. (in press). Scalar implicatures, polarity phenomena and the syntax/pragmatics interface. In A. Belletti (Ed.), *Structures and Beyond*. Oxford University Press
- Chierchia, G., Crain, S., Guasti, M., Gualmini, A., & Meroni, L. (2001). The acquisition of disjunction: evidence for a grammatical view of scalar implicature. In A. H. J. Do (Ed.), *BUCLD Proceedings* (Vol. 25, pp. 157-168). Somerville, MA: Cascadilla Press.
- Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12(4), 335-359.
- Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, 3, 472-517.
- Evans, J. S. B., & Newstead, S. E. (1980). A study of disjunctive reasoning. *Psychological Research*, 41(4), 373-388.
- Fillenbaum, S. (1974). Or: Some uses. *Journal of Experimental Psychology*, 103(5), 913-921.
- Grice, H. P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.
- Horn, L. R. (1973). *Greek Grice: A brief survey of proto-conversational rules in the History of Logic*. Proceedings of the Proceedings of the Ninth Regional Meeting of the Chicago Linguistic Society. 205-214. Chicago.
- Horn, L. R. (1984). Toward a new taxonomy for scalar inference, in D. Schiffrin (ed) GURT. Washington D. C.: Georgetown University Press.

- Howell, D. C. (1997). *Statistical methods for psychology*. Wadsworth: Belmont, CA. 4th edition.
- Just, M. A., & Carpenter, P. A. (1971). Comprehension of negation with quantification. *Journal of Verbal Learning and Verbal Behavior*, 10, 244-253.
- Lea, R. B. (1995). On-line evidence for elaborative logical inferences in text. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 21(6), 1469-1482.
- Lea, R. B., & Mulligan, E. J. (2002). The effect of negation on deductive inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(2), 303-317.
- Lea, R. B., O'Brien, D. P., Fisch, S. M., Noveck, I. A., & et al. (1990). Predicting propositional logic inferences in text comprehension. *Journal of Memory and Language*, 29(3), 361-387.
- Levinson, S. (1983). *Pragmatics*. Cambridge: Cambridge University Press.
- Levinson, S. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. MIT Press: Cambridge, Mass.
- Meyer, D. E. (1970). On the representation and retrieval of stored semantic information. *Cognitive Psychology*, 1, 242-299.
- Moxey, L. M., Sanford, A. J., & Dawydiak, E. J. (2001). Denials as controllers of negative quantifier focus. *Journal of Memory and Language*, 44(3), 427-442.
- Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, 78(2), 165-188.
- Noveck, I. A., Chierchia, G., Chevaux, F., Guelminger, R., & Sylvestre, E. (2002). Linguistic-pragmatic factors in interpreting disjunctions. *Thinking and Reasoning*, 8(4), 297-326.

- Paterson, K. B., Sanford, A. J., Moxey, L. M., & Dawydiak, E. (1998). Quantifier polarity and referential focus during reading. *Journal of Memory and Language*, 39(2), 290-306.
- Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*(10), 437-442.
- Rips, L. J. (1975). Quantification and semantic memory. *Cognitive Psychology*, 7(3), 307-340.
- Sanford, A. J., Moxey, L. M., & Paterson, K. B. (1996). Attentional focusing with quantifiers in production and comprehension. *Memory and Cognition*, 24(2), 144-155.
- Sperber, D., & Wilson, D. (1985/1995). *Relevance: Communication and Cognition*. Oxford: Basil Blackwell.

Author Note

Support for the majority of this work came from by a post-doctoral grant from the *Centre National de la Recherche Scientifique* (France) to the first author as part of an *Action Thematique et Incitative* grant awarded to the second author. The first author is presently supported by NIMH Grant 41704, awarded to Professor G. L. Murphy of New York University. Versions of this paper have been presented at the First International Workshop on Current Research in the Semantics-Pragmatics Interface (Michigan State University, 2003). The authors wish to express their gratitude to Dan Sperber, Jean-Baptiste van der Henst, Nausicaa Pouscoulous, Gregory Murphy, Jennifer Wiley and three anonymous reviewers whose comments improved the paper.

Correspondence concerning this article should be addressed to Lewis Bott, NYU-Psychology, 6 Washington Place, 8th Floor, New York, NY 10003, Lewis.Bott@nyu.edu.

Tables

Table 1

Examples of the Sentence Types used in Experiments 1-4

Reference	Example sentence	Appropriate Response
T1	Some elephants are mammals	?
T2	Some mammals are elephants	T
T3	Some elephants are insects	F
T4	All elephants are mammals	T
T5	All mammals are elephants	F
T6	All elephants are insects	F

Note. T1 sentences are the underinformative sentences referred to in the text. The question mark in the Correct Response column indicates that T1 can be considered true or false depending on whether the participant draws the inference or not.

Table 2

Proportion Responding “True” to Each of the Sentence Types in Experiment 3

Sentence	Example	Mean Proportion True
T1	Some elephants are mammals	0.407 (0.120)
T2	Some mammals are elephants	0.887 (0.018)
T3	Some elephants are insects	0.073 (0.012)
T4	All elephants are mammals	0.871 (0.021)
T5	All mammals are elephants	0.031 (0.006)
T6	All elephants are insects	0.083 (0.017)

Note. Scores are based on N = 32 participants where each participant was required to evaluate 9 instances of each type of sentence. Outlier responses are not included. Variance is shown in parenthesis.

Table 3

Summary of results for Experiment 4

Sentence	Example	Short Lag	Long lag	Response difference
T1	Some elephants are mammals	0.72 (0.053)	0.56 (0.095)	-0.16
T2	Some mammals are elephants	0.79 (0.021)	0.79 (0.038)	0.00
T3	Some elephants are insects	0.12 (0.012)	0.09 (0.007)	+0.03
T4	All elephants are mammals	0.75 (0.027)	0.82 (0.024)	+0.07
T5	All mammals are elephants	0.25 (0.061)	0.16 (0.022)	+0.09
T6	All elephants are insects	0.19 (0.017)	0.12 (0.011)	+0.07

Note. Scores are based on N = 45 participants where each participant was required to evaluate 9 instances of each type of sentence. Outlier responses are not included. The Short lag and Long lag columns contain the proportion of True responses for each condition. Variance is shown in parenthesis. The final column refers to the increase in consistency of responses with added response time. For control sentences this equates to the increase in proportion correct with more time, while for the T1 sentences the figure is the Long condition True response minus the Short condition True response.

Figure Captions

Figure 1. The mean choice proportions and reaction times for Experiment 1. Data is shown as a function of sentence type and instructions given to participants. Error bars refer to the standard error of the mean for the relevant cell of the design.

Figure 2. The mean choice proportions and reaction times for Experiment 2, Agree responses. Data is shown as a function of sentence type and instructions given to participants. Error bars refer to the standard error of the mean for the relevant cell of the design.

Figure 3. The mean reaction times for Experiment 3 as a function of sentence type. Responses to T1 sentences are divided into logical (“true”) and pragmatic (“false”). Error bars refer to the standard error of the mean for the relevant cell of the design.

Figure 1.

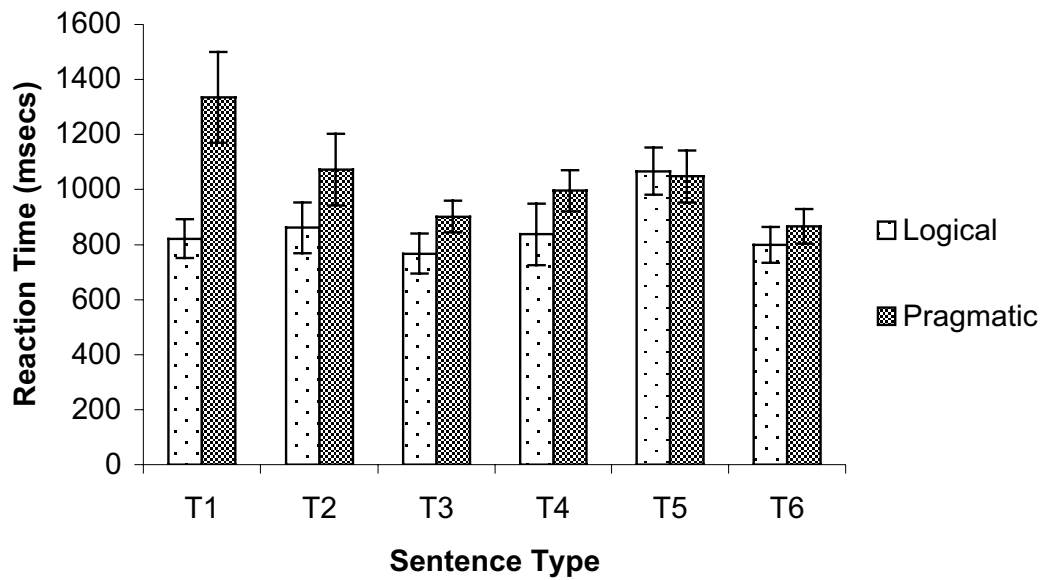
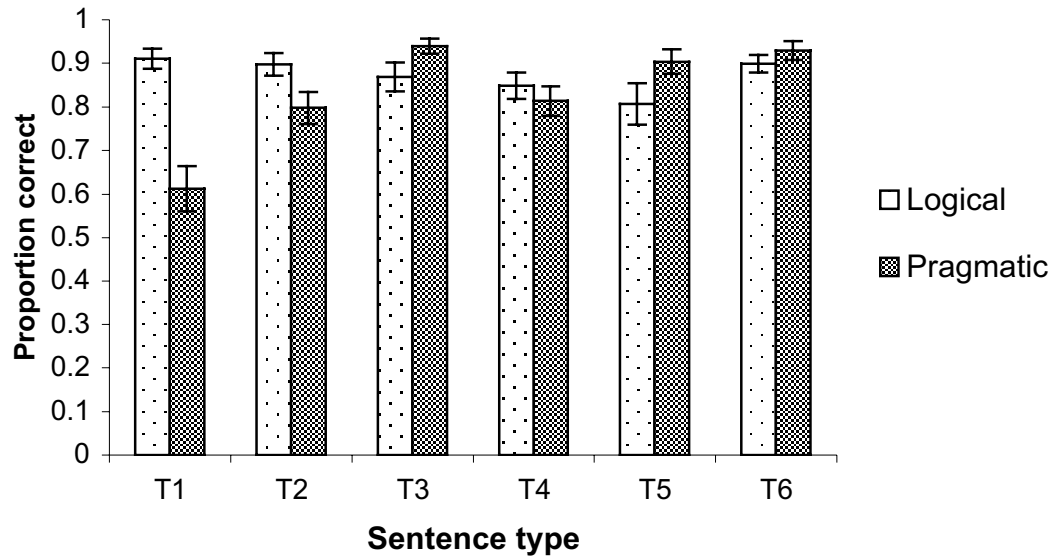


Figure 2

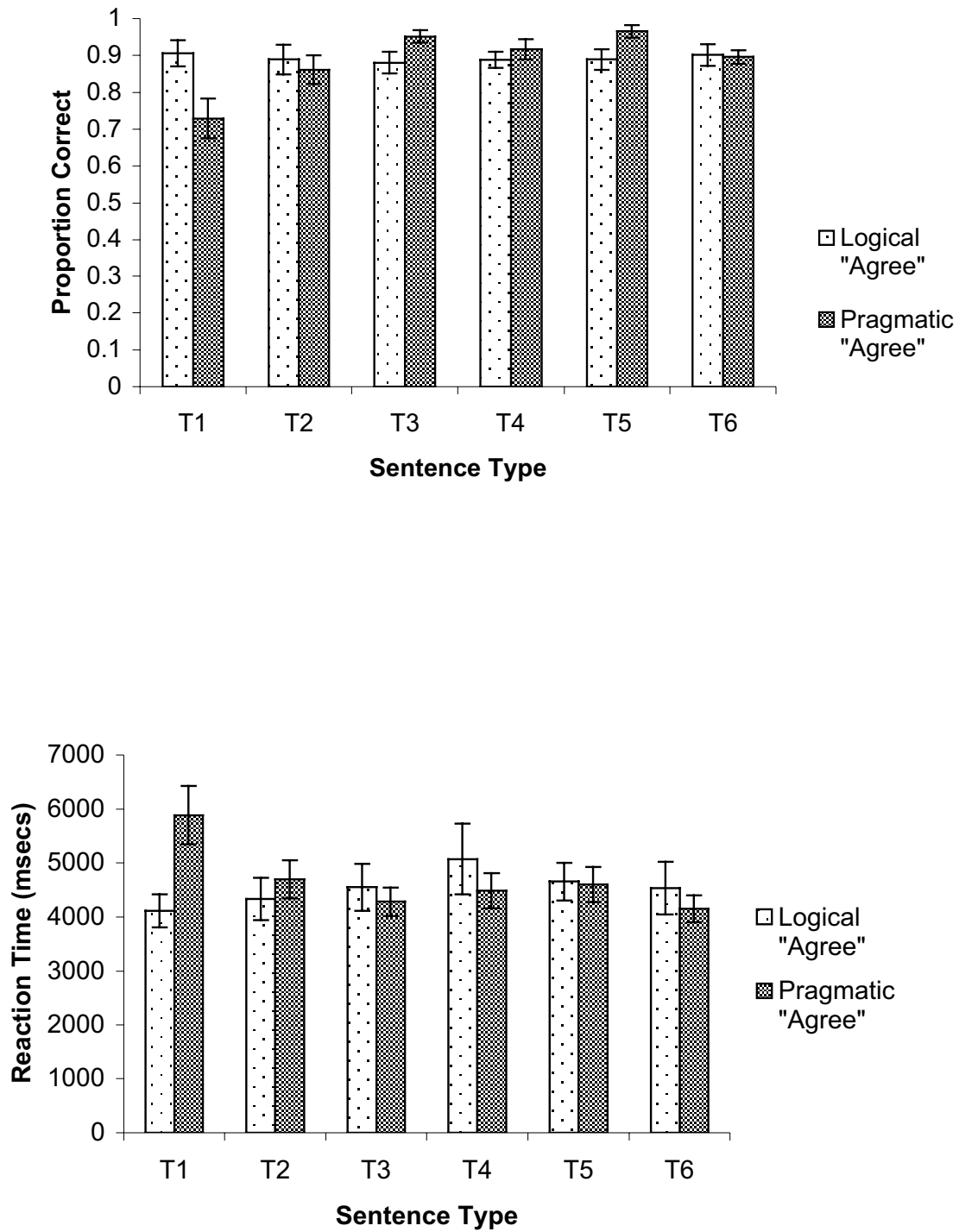
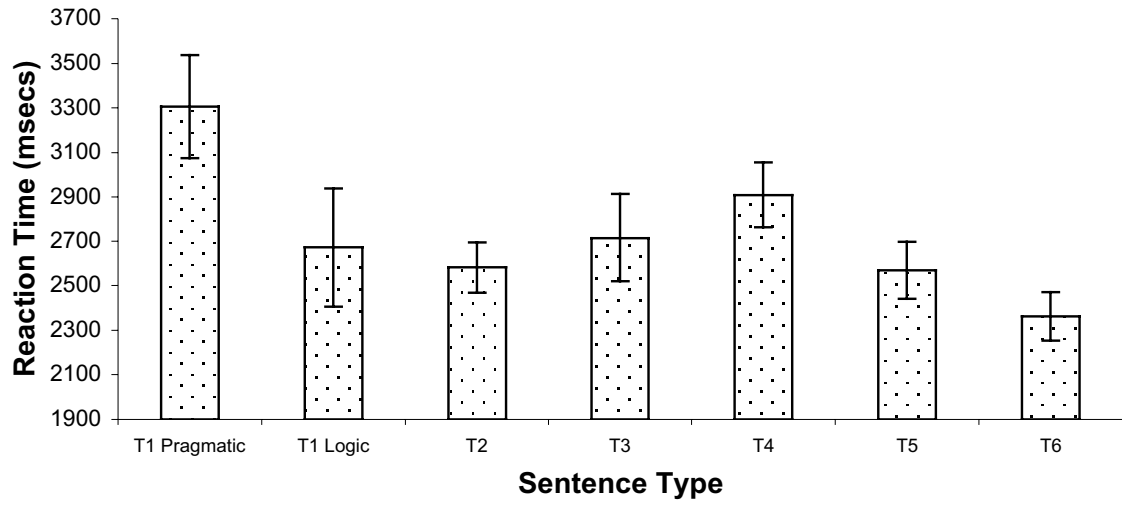


Figure 3



Appendix

Table A.

Categories and Exemplars used in Experiments 1-4

Fish	Reptile	Bird	Mammal	Insect	Shellfish
Anchovies	Alligator	Eagle	Cat	Wasp	Winkle
Carp	Crocodile	Canary	Horse	Spider	Crab
Cod	Frog	Crow	Dog	Cockroach	Prawn
Haddock	Iguana	Owl	Pig	Caterpillar	Clam
Piranha	Lizard	Sparrow	Elephant	Ant	Oyster
Shark	Salamander	Peacock	Sheep	Fly	Lobster
Salmon	Snake	Parrot	Bear	Mosquito	Langoustine
Tuna	Tortoise	Pigeon	Monkey	Butterfly	Mussel
Trout	Newt	Vulture	Cow	Beetle	Cockle

Note. The categories are shown in the top row while exemplars of each category are shown in the corresponding column. Stimuli are translated from French.