Interpreting numerals and scalar items under memory load

Paul Marty & Emmanuel Chemla & Benjamin Spector
LSCP-IJN (CNRS, EHESS; DEC, ENS, Paris, France)

February 4, 2012

Abstract

A sentence such as 'Four children are blond' can be interpreted as meaning either that *at least four* children are blond (weak reading), or that *exactly four* children are blond (strong reading). On the classical neo-Gricean view (Horn 1972), this ambiguity is similar to the ambiguity generated by scalar terms such as 'some', for which both a weak reading (i.e., *some (possibly all)*) and a strong reading (i.e., *some but not all*) are available. On this view, the strong reading of numerals, just like the strong reading of 'some', is derived as a scalar implicature (SI), on the basis of the weak reading. However, more recent studies have found significant differences between the two phenomena. The syntactic distribution of the strong reading is not the same in both cases (Horn 1992, Breheny 2008), and children seem to acquire the strong reading of numerals before they acquire the strong reading of standard scalar items (Noveck 2001, a.o.). Using a dual task approach, we provide evidence for another type of difference between numerals and standard scalar items. We show that tapping memory resources has opposite effects on 'some' and on bare numerals. Under high cognitive load, participants reported fewer SIs for sentences involving 'some' (compared to low cognitive load conditions), but they report *more* strong readings for sentences involving bare numerals. We discuss the consequences of this result for the current theoretical debates regarding the semantics of numerals.

Keywords: numerals; scalar implicatures; pragmatics; working memory; dual-task.

Introduction

It is well known that bare numerals (*one*, *two*, *three*, ...) give rise to a systematic ambiguity between an 'exact' reading and an 'at least' reading. Thus a sentence such as (1) below can be understood as conveying either that John has exactly three children, or that he has at least three, depending on the context in which the sentence is uttered:

(1) John has three children.

To illustrate, consider the two following dialogues:

(2) a. How many children does John have?
b. He has three children.

(3) a. In this country, one needs to have three children in order to qualify for a tax deduction. What about John? Does he qualify?
b. Yes, he does. He has three children. In fact, he even has four.

While the sentence 'He has three children' is preferably interpreted, in the context of (2), as conveying that John has three children and no more, the very same sentence receives an 'at least'-interpretation in the context of (3).

In order to account for this ambiguity, several types of accounts have been proposed in the literature. The most traditional one, which we label the 'neo-Gricean approach', assumes that numerals are not ambiguous in terms of their *linguistic* meaning, and that the literal, linguistic meaning of (1) corresponds to the 'at least'-interpretation (Horn 1972). According to this approach, the 'exact'-interpretation is the result of a pragmatic reasoning which takes as input the literal 'at least'-meaning and returns a *strengthened* meaning corresponding to the 'exact'-interpretation. This pragmatic strengthening is viewed as a special case of a more general phenomenon, namely, the phenomenon of *scalar implicature* whereby, e.g., a sentence such as 'Mary solved some of the problems' is understood to imply that Mary didn't solve all the problems, even though this sentence, under its literal interpretation, is true if Mary solved all problems.

In a nutshell, the neo-Gricean approach says the following. Suppose someone utters the sentence 'John has three children'. This sentence means that John has at least three children. However, if the speaker had believed that John had more than three children, she could have said so, for instance by saying 'John has four children', and in doing so she would have been more cooperative (Grice's maxim of *quantity*). Hence, if the speaker is cooperative, she does not believe that John has more than three children. Since the speaker knows that I am going to draw this inference, she is entitled to consider that I will interpret her utterance as conveying that John has exactly three children. This pragmatic derivation of the 'exact'-reading for numerals is entirely parallel to the derivation of the some-but-not-all reading in the case of 'some'. If someone utters 'Mary solved some of the problems', one is entitled to conclude that the speaker does not believe that Mary solved all of the problems, for if she did, she should have said 'Mary solved all of the problems'. In the last decade, the neo-Gricean, pragmatic approach to scalar implicatures has been challenged by advocates of a grammatical theory of scalar implicatures (cf. Chierchia et al. to appear and the references cited therein). Importantly, these approaches generally also assume that the 'exact'-interpretation of numerals is an instance of a scalar implicature.

Recently, the view that the 'exact' reading of numerals is to be derived as a scalar implicature has been questioned, on various grounds. First, numerals can give rise to an 'exact' reading even in syntactic environments where standard scalar items such as 'some' tend not to be interpreted according to their 'strengthened', upper-bounded meaning ('some but not all'), such as negative and other downward-entailing environments (cf., e.g., Horn 1992, Breheny 2008). Second, several experimental studies (Noveck 2001, Papafragou & Musolino 2003, Musolino 2004, Guasti et al. 2005, Pouscoulous et al. 2007, Huang & Snedeker 2009b) have shown that young children tend to compute 'exact' readings for numerals to a much greater extent than they compute standard scalar implicatures. On the basis of these ob-

servations, several recent works argue, *contra* the traditional neo-Gricean approach, that the primary linguistic meaning of numerals corresponds to the 'exact' reading (see, e.g., Geurts 2006, Breheny 2008).[1] However, the facts regarding the syntactic distribution and the acquisition of the 'exact' reading do not in themselves exclude the possibility that the 'exact' reading of numerals is a kind of scalar implicature. First, even standard scalar items can sometimes retain their strengthened meaning in negative and other downward-entailing contexts, in particular if they bear focal stress (cf., e.g., Levinson 2000). It is conceivable that numerals are different from other scalar items in that they are lexically focused, which could explain both why the 'exact' reading is easily accessible in negative contexts and why it seems to be particularly salient in non-negative contexts (scalar items yield more robust scalar implicatures when focused, cf. Zondervan 2010). Second, the developmental observations we have just mentioned would also be expected if the scale of numerals were acquired before other scales, for instance because it is cognitively much more salient (a possibility that squares well with the idea that numerals are intrinsically focused). In fact, several experimental works provide evidence that young children's performance in tasks involving scalar implicature computation significantly improves when the relevant alternatives are explicitly present in the task or are made contextually salient (see, e.g., Chierchia et al. 2001, Barner et al. 2011). While it is clear that numerals behave differently from standard scalar items, the hypotheses we have just sketched can explain these differences without forcing us to reject the view that the 'exact' reading of numerals is derived as a scalar implicature.

The present work aims at contributing to this debate by bringing a new type of evidence. Our goal is to compare the *processing costs* involved in the derivation of the relevant readings for both scalar items such as 'some' and bare numerals. Our results will reveal a new difference between the two cases. In the case of scalar items, previous studies have shown that the derivation of the stronger reading (e.g., 'some but not all' in the case of 'some') involves more effort than the derivation of the plain, literal reading ('some and possibly all'): in studies involving response time (Noveck & Posada 2003, Bott & Noveck 2004, Breheny et al. 2006, Huang & Snedeker 2009a), subjects were found to take more time to compute the stronger reading than the weaker reading, and in recent dual-task studies (De Neys & Schaeken 2007, Marty & Chemla 2011), the stronger reading attached to scalar items was shown to come at an extra memory cost.

In a dual-task experiment, as we will explain below, participants are asked to perform two tasks simultaneously, and both tasks are assumed to compete for the same processing resources. If the first task involves the derivation of a scalar implicature, and the second task is a short-term memory task, then a decrease in performance in the first task due to the complexity of the second task can be interpreted as showing that short-term memory is involved in the derivation of scalar implicatures. No such study has been undertaken in the case of numerals. We present the results of a dual-task experiment which involves both the scalar item 'some' (or rather, its French counterpart 'certains') and bare numerals. These results will show that computing the *strong*, 'exact' reading of numerals is *less* demand-

[1]For Geurts (2006), numerals are ambiguous between an 'exact' interpretation and an 'at least' interpretation, but the 'at least' interpretation is derived from the 'exact' reading by means of two successive type-shifting operations. Breheny (2008) offers a theory which is the mirror-image of the neo-Gricean approach to numerals: according to him, 'at least' interpretations are the result of a special pragmatic mechanism which takes as input the 'exact' interpretation and *weakens* it into an at-least interpretation.

ing, in terms of short-term memory, than computing the *weak*, 'at least' reading. In the case of 'some', our results replicate previous experimental studies that showed an opposite pattern. As we will show, this contrast, and the other differences that previous studies have established, can receive quite different accounts. It can be explained by assuming that the 'exact' reading corresponds to the primary linguistic meaning of numerals (following previous works), but we will point out that there exist other plausible explanations.

Before going on, an important clarification is in order. As mentioned above, in recent years, the pragmatic approach to scalar implicatures has been challenged by so-called 'grammatical' theories of scalar implicatures (e.g., Chierchia 2004, Chierchia et al. to appear). In such theories, the derivation of the 'some-but-not-all' reading for 'some' is not viewed as a pragmatic inference, but as involving a specific grammatical mechanism. This paper does not directly address the debate between pragmatic and grammatical approaches to scalar implicatures. Our main goal is to compare scalar items such as 'some' and numerals, in order to determine whether the derivation of the 'stronger' readings involves the same mechanisms for both cases. Whatever the answer to this question is, it is in large part neutral with respect to the debate between pragmatic and grammatical approaches to scalar implicatures, even though it will certainly constrain in various ways the precise shape that such theories will take. We will point out the more general theoretical consequences of our findings in the discussion.

Experiment

We used a dual-task paradigm to compare the working memory cost involved in the derivation of SIs and in the computation of 'exact', strong readings of numerals. Participants were asked to perform a graded sentence-picture matching task (Chemla & Spector 2011) on SI-sentences (e.g., *Some dots are red*) and Numeral-sentences (e.g., *Four dots are red*), while they simultaneously had to memorize a sequence of letters (see Figure 2).

The central executive component of working memory is assumed to be responsible for coordination of the various processes involved in short-term storage tasks and dual-task coordination is widely considered as one of its main functions (Baddeley 1992, Miyake & Shah 1999, Engle et al. 1999). In the present study, the cognitive load on working memory was manipulated by varying the length of the sequence of letters to be memorized so that participants' memory resources were either minimally busy or more heavily tapped during the linguistic task.

The rationale for this dual task procedure is that participants should not be able to appeal to their working memory for the linguistic task in conditions where the cognitive load is high, i.e. in conditions where working memory is needed to perform the memory task. Hence, any process that also requires these resources should be impaired. Our results show that 'exact'-interpretations of numerals and strong readings associated to scalar items such as 'some' are not impacted in the same way by this procedure.

Participants

The participants were 26 native speakers of French, aged between 19-34 years (mean age 23 years; 13 females). All of them reported to have normal color vision and no prior exposure to formal linguistics.

Materials and Tasks

Letter memory task. The memory task was a classical storage task of letters. Input sequences consisted of 2 or 4 capital letters. Each letter was presented in the center of the screen for 1 second, with a 500 ms blank screen in between letters. The nine letters used were B, F, H, J, L, M, Q, R, and X, chosen to be phonologically dissimilar. Participants first memorized the sequence. After the truth value judgment task (see below), they were asked to type in the sequence of letters in backward order.

It is known that the cognitive effort required to encode such a sequence depends on several factors such as, for instance, the number of elements in the sequence (Baddeley et al. 1975) or the phonological similarity between these elements (Conrad & Hull 1964, Baddeley 1966). For the present study, the complexity of the to-be-memorized sequence was manipulated only by varying the number of its elements to obtain 2-letter and 4-letter sequences, respectively used in the LOW LOAD and in the HIGH LOAD trials. The working memory resources should be minimally burdened by the 2-letter sequences, and more so by the 4-letter sequences.

The load factor was manipulated within subject. Participants were administered two consecutive blocks of trials: one block contained LOW LOAD trials and the other block contained HIGH LOAD trials. For each participant, it was pseudo-randomly determined which type of block they started with.

Sentence-picture matching task. Each experimental item consisted of a sentence and a picture. These two components were displayed on the screen as in the examples given in Figure 1.

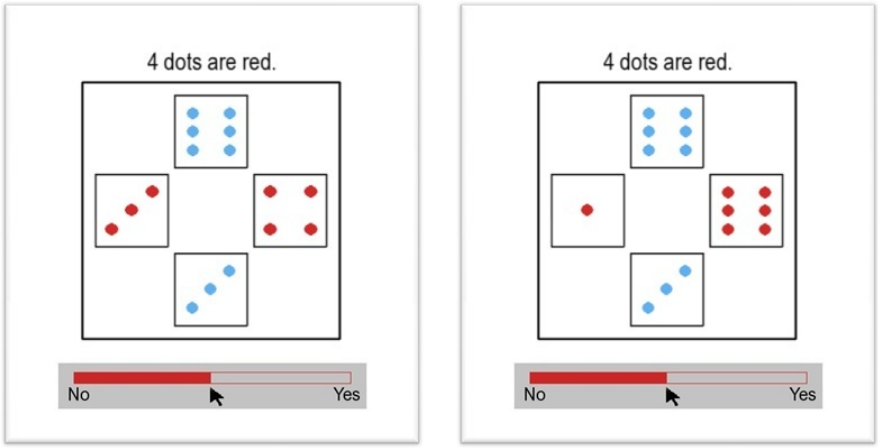


Figure 1. Examples of the sentence-picture display.

For each sentence-picture item, participants were asked to assess the extent to which the sentence was a correct description of the situation represented on the picture. They gave their answers along a continuum of answers, by setting with a cursor the right end of a red line going from 'No' (to the left) to 'Yes' (to the right). Answers were coded as the

position of the response on the scale, from 0% for a rejection and 100% for acceptance of the sentence as a correct description (see Chemla & Spector 2011).

Experimental items. All the sentences were of the form '⟨Quantifier⟩ dots are ⟨color⟩', where ⟨Quantifier⟩ varies as described in Table 1. The ⟨color⟩ was one of the following color adjectives: 'rouge'(*red*), 'vert'(*green*) or 'bleu'(*blue*). These colors were chosen to be easily identifiable. The value of n used for the numeral quantifiers was either 3 or 4.

Name	Description of the sentence	
Some	Some	dots are ⟨color⟩.
All	All	dots are ⟨color⟩.
Bare	$(n+1)$	dots are ⟨color⟩.
Between	Between n and $(n+2)$	dots are ⟨color⟩.
More than	More than $(n-1)$	dots are ⟨color⟩.
Fewer than	Fewer than $(n+3)$	dots are ⟨color⟩.

Table 1: Schematic description of the sentence types used in the sentence-picture matching task. For a more concrete illustration, you may read n as *3* and ⟨color⟩ as *red*.

All the pictures were composed of four squares. Each square contained between 1 and 6 dots, which were represented as on the faces of a dice, so that counting and adding the number of dots would be easy. In each square, dots were either of the target color used in the sentence (abbreviated to target dots henceforth), or of a different filler color (red, blue, green, purple, yellow, black or gray).[2] The number of target dots represented on the pictures varied as described in Table 2.

Name	Description of the picture	
Null	No	dots are ⟨color⟩.
Partial	Only some	dots are ⟨color⟩.
Total	All	dots are ⟨color⟩.
Inferior	$(n-2)$	dots are ⟨color⟩.
Intermediate	$(n+1)$	dots are ⟨color⟩.
Superior	$(n+4)$	dots are ⟨color⟩.

Table 2: Schematic description of the picture types used in the sentence-picture matching task, where n and ⟨color⟩ refer respectively to the value of n and the color adjective involved in the sentence they were paired with.

The full list of conditions is given in Table 3. The number of repetitions is given in parenthesis for each condition. The cases or primary interest involved the *Some* (4), *Bare* (5) and *Between* (6) sentences, which are ambiguous between a weak reading and a strong reading, as shown in the following examples. Although we were mostly interested

[2]The ⟨color⟩ involved in each sentence-picture item was randomly selected from the list of target colors. The second color used in the picture was then pseudo-randomly chosen from the list of filler colors minus the selected target color.

in the comparison between genuine cases of scalar implicatures and cases involving bare numerals, we initially set our study to investigate another possible ambiguity of the same type involving the phrase 'Between n and m'. This new ambiguity phenomenon has been recently studied in Marty et al. (2012). We will not put any strong emphasis on these cases as the results are harder to interpret (these ambiguities behave somewhat intermediately between the other two).

(4) Some dots are red.
   a. Weak reading: Some dots are red (possibly all).
   b. Strong reading: Some dots are red, but not all.

(5) 4 dots are red.
   a. Weak reading: At least 4 dots are red.
   b. Strong reading: Exactly 4 dots are red.

(6) Between 3 and 5 dots are red.
   a. Weak reading: At least 3 dots are red.
   b. Strong reading: At least 3 and at most 5 dots are red.

In the critical conditions, these sentences were paired with pictures that make them true under their weak reading, but false under their strong reading (**Target** conditions). In the control conditions, they were paired with pictures that make them either true or false under both of their readings simultaneously (**True** and **False** conditions).

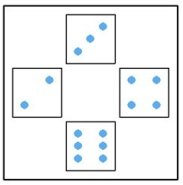
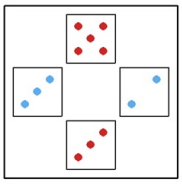
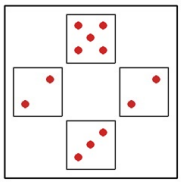
We added further controls in which participants were asked to judge unambiguous sentences. *All* sentences were included to control that participants have no problem understanding the (negation of the) *not-all* proposition involved in the SI derivation for the *Some* sentences.[3] *More than* and *Fewer than* sentences were included to ensure that participants could count the dots properly and do the task appropriately. These control sentences were paired with pictures that make them either true or false (**True** and **False** conditions).

Nota bene. For the sentences involving a numeral quantifier, we also manipulated the grouping of target dots represented on the pictures they were paired with. Our goal was to test whether participant's judgment would differ in situations where a particular group of target dots makes the sentence true when the overall picture does not. For instance, a sentence like '4 dots are red' was paired in the group situations with pictures on which a group of 4 red dots was easily identifiable (see example on the left in Figure 1). Conversely, no such group was available in the no-group situations (see example on the right in Figure 1). In the rest of the paper, we will set aside the group/no-group distinction as results showed that there was no difference in participants' responses between the two kinds of situations. It confirms that participants fully understood the task and considered each picture as a whole when judging the sentences.

Procedure

Each trial started with the presentation of a sequence of letters to be memorized. Then, a sentence-picture item was displayed on the computer screen and remained until

[3]Note that the alternative sentence in a SI theory of bare numerals is again a bare numeral sentence, which was thus automatically included in the experiment.

Sentence \ Picture	*Null* 	*Partial* 	*Total* 
Some Some dots are red.	**False** (4)	**True** (4)	**Target** (8)
All All dots are red.	—	**False** (4)	**True** (4)

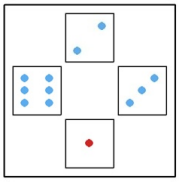
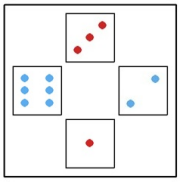
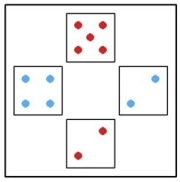
Sentence \ Picture	*Inferior* 	*Intermediate* 	*Superior* 
Bare 4 dots are red.	**False** (8)	**True** (8)	**Target** (8)
Between Between 3 and 5 dots are red.	**False** (8)	**True** (8)	**Target** (8)
More than More than 2 dots are red.	**False** (4)	—	**True** (8)
Fewer than Fewer than 6 dots are red.	**True** (4)	—	**False** (8)

Table 3: Summary of the sentence-picture matchings giving rise to the **Target**, **True** and **False** conditions (with $n = 3$ and *red* as the target color for these examples). Numbers in parenthesis refer to the number of test items included in the experiment to exemplify the different conditions.

the participant provided a response. Next, participants had to reproduce the sequence of letters in backward order. At the end of each trial, they received feedback on the accuracy of the reproduction. A depiction of the general structure of a trial is given in Figure 2.

Participants were instructed that it was crucial for the experiment to reproduce accurately the complete sequences of letters. They were also encouraged to use the flexibility of the red line to represent at best their intuition concerning the correspondence between the sentence and the situation represented on the picture.

Participants started with a short training composed of 4 complete trials (2 with 2-letter sequences and 2 with 4-letter sequences). The sentences used for the training were unrelated to the present experimental issue and were simply included to help participants familiarize with the display (e.g., *There are red dots*). Participants were then given two blocks of 96 sentence-picture items (see Table 3 for details) with a short break in between.

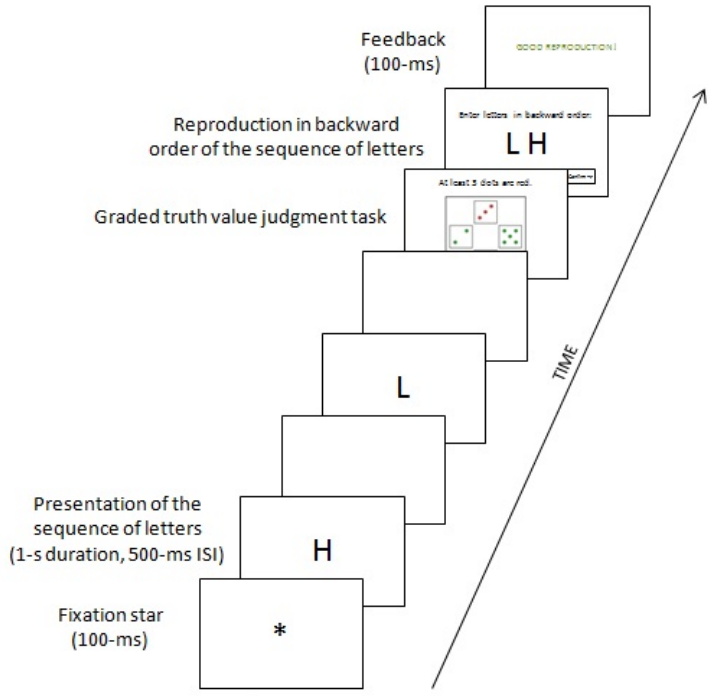


Figure 2. Dual-task procedure: general structure of a trial.

The proportion of expected True/False responses was balanced within as well as between blocks. In each block, test items were presented in random order.

Experimental Hypothesis

Previous experimental studies (De Neys & Schaeken 2007, Marty & Chemla 2011) have shown that some of the cognitive processes involved in the computation of SIs draw on executive working memory resources. More precisely, it was found that participants derive fewer SIs when they are under high cognitive load. If the strong reading of numerals is derived through the same process, it should involve executive memory resources in the same way. Hence, in the critical cases in which the strong reading is false and the weak reading is true, participants should (i) judge the *Some* sentences more appropriate (fewer SIs) in the HIGH LOAD than in the LOW LOAD trials, (ii) they should similarly judge the *Bare* sentences more appropriate (fewer strong readings) in the HIGH LOAD than in the LOW LOAD trials.

Results

Letter Memory Task

For the letter memory task, a response was coded as correct if the participant reproduced the sequence of letters correctly in its entirety. The sequences of letters were reproduced accurately in 94% ($SD = 5$) of the trials, with a better performance for the LOW LOAD than for the HIGH LOAD trials: $M = 96\%$ *vs.* $M = 92\%$, $t(25) = 4.17$, $p < .001$.

These results show that the letter memory task was properly performed and confirmed that the 4-letter sequences were more demanding than the 2-letter sequences.

Data Treatment

In all subsequent analyses, items for which participants did not accurately reproduce the complete sequence were removed (about 6% of the trials). We also removed the items with the 2.5% fastest and 2.5% slowest response times in the linguistic task (about 5.5% of the responses). According to a monofactorial analysis of variance (ANOVA), the average rate of removed trials did not significantly differ from one sentence type to another in LOW LOAD as well as in HIGH LOAD trials (Fs < 2, n.s.).

Sentence-Picture Matching Task

Control sentences. Mean responses for the *All*, *More than* and *Fewer than* sentences are reported in Figure 3.

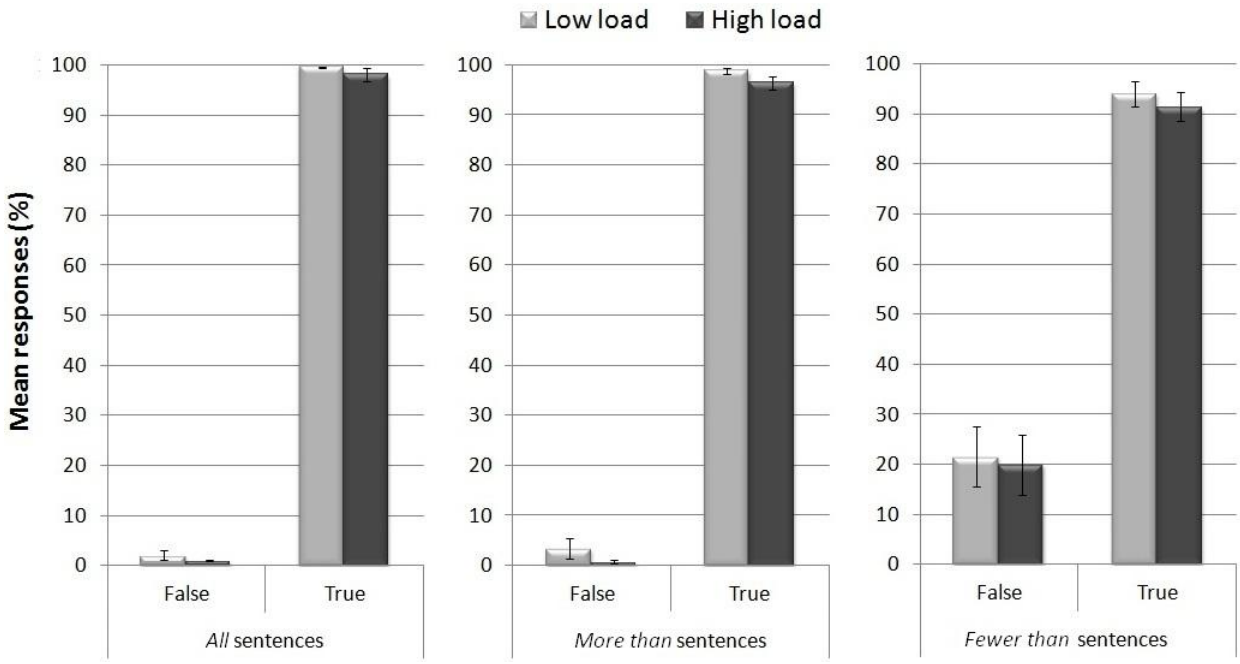


Figure 3. Mean responses (%) for the control sentences as a function of conditions (**True** vs **False**) in LOW LOAD and HIGH LOAD trials. Error bars refer to standard errors.

Participants' performance for the control sentences were overall as expected. These sentences received a global mean score (i.e. across LOW LOAD and HIGH LOAD trials) of 8% ($SD = 13$) in the **False** condition and of 96% ($SD = 8$) in the **True** condition.[4]

[4]The global mean score for the *Fewer than* sentences in the **False** condition is somewhat higher than the score obtained for the *More than* sentences in its own **False** condition. There exist plausible explanations for this discrepancy. As shown in Geurts & Der Slik (2005), downward-entailing quantifiers like 'fewer than' are more difficult to process than upward-entailing quantifiers like 'more than'. The relative failure with the former in our task may reflect the same effect. More tentatively, this result could indicate that an existential,

We conducted paired-samples t-tests to compare participants' responses in the LOW LOAD and the HIGH LOAD trials and none of these tests reached significance (all ts $<$ 1.7, n.s.).

In sum, these results show that participants performed the task appropriately, and that the concurrent memorization of the 4-letter sequences did not interfere with their understanding of unambiguous sentences.

Target sentences. Mean responses for the *Some*, *Bare* and *Between* sentences are reported in Figure 4.

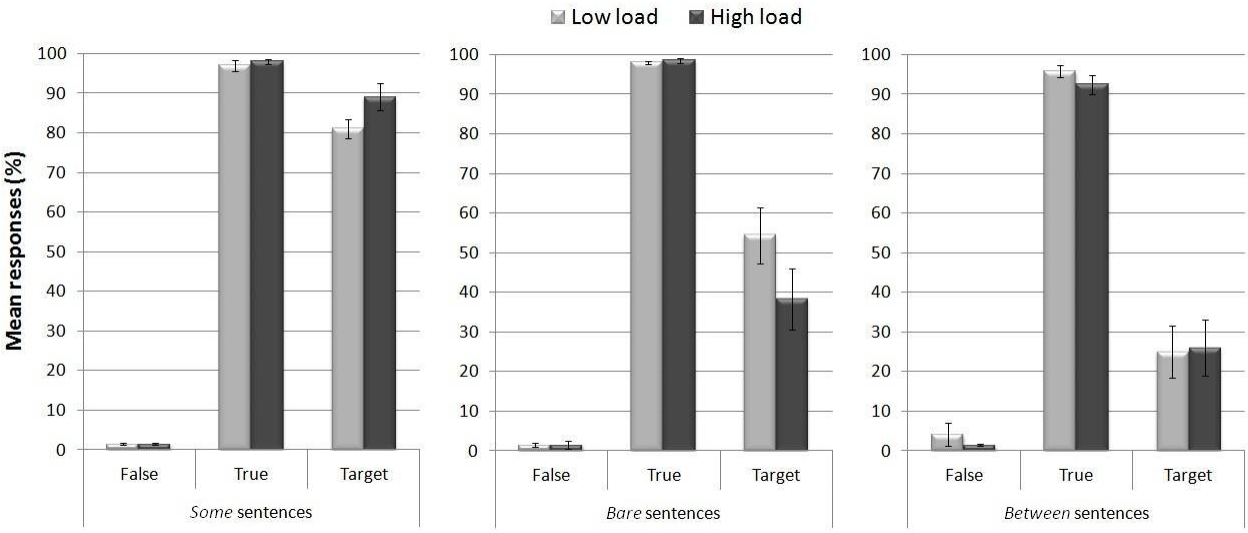


Figure 4. Mean responses (%) for the target sentences as a function of conditions in LOW LOAD and HIGH LOAD trials. Error bars refer to standard errors.

In the control cases, participants' performance for the target sentences was as expected. They received a global mean score of 1% ($SD = 4$) in the **False** condition and of 96% ($SD = 6$) in the **True** condition. There was no difference between HIGH and LOW LOAD trials (all ts < 1.2, n.s.).

In the **Target** conditions, mean responses were all found to be significantly higher than in the **False** condition (all ts > 3.5, ps $< .001$), and lower than in the **True** condition (all ts > 5, ps $< .0001$). These results confirm that target sentences are ambiguous between a weak reading and a strong reading.

The crucial finding is that the dual task had an opposite memory effect on the interpretation of the *Some* and *Bare* sentences in the target condition. *Some* sentences were judged significantly more appropriate in the HIGH LOAD than in the LOW LOAD trials ($M = 89\%$ *vs.* $M = 81\%$, $t(25) = 2.68$, $p < .01$), whereas *Bare* sentences were judged significantly *less* appropriate ($M = 38\%$ *vs.* $M = 54\%$, $t(25) = 4.33$, $p < .001$). The

weak reading is also marginally available for the *Fewer than* sentences. On this reading, a sentence such as 'Fewer than four dots are red' would mean that there exists a plurality X made up of fewer than four red dots. This is in fact true as soon as there is at least one red dot, i.e. no matter how many red dots there are. If there are ten red dots, then one can find a plurality made up of fewer than four red dots.

interaction between type of sentence (Some *vs.* Bare) and type of load (Low *vs.* High) in the **Target** condition was significant: $F(1, 25) = 25.9,\ p < .0001$.

Hence, participants made fewer strong readings (with SIs) for the *Some* sentences when their executive resources were tapped, but *more* strong readings for the *Bare* sentences. Overall, these findings show that the processing of strong readings associated to 'some' and to bare numerals does not recruit in the same way the central component of working memory.

Discussion

First, our findings regarding scalar items such as 'some' replicate previous results (De Neys & Schaeken 2007, Marty & Chemla 2011): the strong reading (i.e. with an SI) attached to the scalar term 'some' comes at a memory cost. Second, our original contribution is that this effect is reversed for numerals: the *weak* ('at-least') reading of numerals is the one that comes at a memory cost, compared with the *strong*, 'exact' reading. More specifically, we found that tapping participants' executive resources induces a significant decrease of SIs for sentences with 'some', whereas it generates a significant increase of strong readings for sentences with bare numerals.

These new findings provide direct evidence for the view that the ambiguities generated by bare numerals and standard scalar items are *processed* differently. These processing differences are in line with other differences that had already been noted (differences regarding the syntactic distribution of the weak and strong readings and developmental differences). These results can be interpreted as going against the traditional view that the strengthened meaning of bare numerals and regular scalar items are obtained through similar processes. In particular, they may suggest that the 'exact' reading of numerals is not derived from the 'at least' reading by way of scalar implicature.

However, we would like also to consider a couple of alternative explanations of our data. Let us assume a scalar implicature view of numerals. There remain a number of possible steps in the derivation of the strong, 'exact' reading that may create behavioral and linguistic differences. First, the alternatives involved in the two cases have different properties. Numerals create minimally different alternatives ('four' is the relevant alternative to 'three', 'five' is the relevant alternative to 'four', etc.). In contrast with this, the alternative of 'some' is obtained by a replacement with 'all', which is not as minimally similar to its competitor. Such a distance in the case of 'some' might create a specific memory cost linked to the process of identifying the relevant alternative (see also our remarks in the introduction of this paper about the interpretation of developmental data). However, Marty & Chemla (2011) found that the memory cost found for 'some' disappears with 'only some' constructions, which involve the same alternatives. This additional datapoint suggests that the source of the memory cost cannot be solely located in the derivation of the relevant alternative.

Second, one may argue that our results could be due to a different kind of difference between the strong readings attached to bare numerals and to 'some'. Scalar implicatures might be computed 'by default' in the case of bare numerals, but not in the case of other scalar items. In other words, there could be an irrepressible tendency to compute scalar implicatures associated with bare numerals (see Levinson 2000 for the view that readings that include SIs are *generally* the default readings). As a result, the strong reading would

come at no additional (visible) cost, but there would be a cost to 'undo' this reading in favor of the weak reading. One may thus try to reduce the difference between 'some' and bare numerals to different tendencies to compute the associated scalar implicatures. In any event, although theoretically possible, this proposal remains incomplete, and there is little independent data that we could use to defend this proposal

There is a third aspect of the semantics of bare numerals that may contribute to a difference with 'some'. Even if we assume that the 'at least' reading is the primary reading for sentences involving bare numerals as quantifiers, the way this reading comes about is generally assumed to involve a non-trivial *logical* (as opposed to pragmatic) inference. Given standard assumptions regarding the semantics of numerals, particularly in dynamic semantics (see e.g., Kadmon 1985, 2001, Krifka 1996, Nouwen 2003, Rooij & Schulz 2006), bare numerals give rise to an 'at least' reading only when they combine with distributive predicates, which license certain inference patterns. Here is why. The lexical entry for bare numerals is assumed to be something like this:

(7) $[\![\text{four}]\!] = \lambda P.\lambda Q.\exists x(\#x = 4 \wedge P(x) \wedge Q(x))$ (where $\#x = 4$ means that x is a plural individual that contains exactly 4 atomic members, assuming standard mereology)

In the case of 'four dots are red', this results in the following truth-conditions:

(8) There exists a plural individual x containing exactly four atomic members, and x qualifies as *dots* and as *red*, i.e. each atomic member of x is a red dot.

Now, it turns out that this is equivalent to the 'at least' reading (i.e. 'at least four dots are red'), for the following reason: suppose that there are more than four red dots, say six. Then one can find a plurality consisting of exactly four objects which are red dots (namely, pick any four dots within the six red dots). The sentence thus comes out true if there are exactly four red dots, but also if there are more than four red dots. But note that this reasoning involves an inferential step, which crucially relies on the fact that *dots* and *red* are distributive predicates, in the following sense: if a certain collection X of objects belongs to the denotation of *dots* (resp. *red*), then any sub-collection of X also belongs to the denotation of *dots* (resp. *red*). So computing the 'at least' reading requires a non-trivial inferential step from groups to subgroups.[5] This inferential step might plausibly involve some memory cost, even if it does not involve any scalar implicature. And this alone could be sufficient to explain why it is easier to identify the sentence 'Four dots are red' as true when coupled with a picture containing exactly four dots than when the picture contains

[5]This reasoning does not go through if a non-distributive predicate is used. Consider for instance the sentence 'Twenty soldiers surrounded the castle'. According to a lexical entry such as the one in (7), this sentence is predicted to mean that there is a group made up of exactly twenty soldiers and that this group surrounded the castle. Assume now that there is a group of thirty soldiers who surrounded the castle, forming together a circle around it, and that removing any ten soldiers would leave one side of the castle without any soldier facing it. Then it is *not* possible to pick a subgroup made up of exactly twenty soldiers such that this subgroup surrounded the castle. This shows that the sentence 'Twenty soldiers surrounded the castle' is not necessarily true whenever more than twenty soldiers surrounded the castle, and therefore this sentence is not expected to have an 'at-least' reading. Likewise, if exactly five people lifted a piano together, it does not follow that there is a group of four people that lifted the piano together, and therefore the sentence 'four people lifted the piano together' is not necessarily true if more than four people lifted the piano together.

more than four dots: in such a case, the inferential step from groups to subgroups is not necessary in order to assess the truth-value of the sentence. Namely, the existence of a plurality made up of exactly four red dots might be immediately noticed in this case.

Conclusion

Bare numerals are often treated as standard scalar items. However, the present results reveal a new kind of difference between the ambiguities generated by bare numerals and by regular scalar items such as 'some': they are impacted differently by external memory demands. This finding adds to the existing body of evidence suggesting that the 'exact' reading of bare numerals does not arise in the same way as the strong readings of scalar items does. In particular, it provides an additional argument for the view that the 'exact' reading of numerals is the primary meaning (as opposed to an implicature-based account of the 'exact' reading).

We note that a careful look at the underlying theories reveal the existence of possible alternative explanations for these differences between bare numerals and other scalar items (and for other differences that have been noted in the literature). In any case, our results create a specific *desideratum* for developing a formally explicit account of the semantics and pragmatics of numerals: any such proposal should be compatible with the fact that the 'exact' reading of numerals is processed differently from the 'strengthened' reading of standard scalar items.

Acknowledgments

We would like to thank Ira Noveck, François Récanati, audiences at LSCP in June 2001 and at MIT in October 2001 for very useful comments on earlier versions of this paper. We are also grateful to Isabelle Brunet and Anne-Caroline Fievet for their invaluable practical help. This work was supported by a 'Euryi' grant from the European Science Foundation ("Presupposition: A Formal Pragmatic Approach") and by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-20013)/ERC grant agreement #229 441-CCC.

Bibliography

Baddeley, A. (1966). The influence of acoustic and semantic similarity on long-term memory for word sequences. *The Quarterly Journal of Experimental Psychology*, *18*(4), 302–309.

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556.

Baddeley, A., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *14*(6), 575–589.

Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition*, *118*(1), 84–93.

Bott, L., & Noveck, I. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of memory and language*, *51*(3), 437–457.

Breheny, R. (2008). A new look at the semantics and pragmatics of numerically quantified noun phrases. *Journal of Semantics*, *25*(2), 93.

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

Chemla, E., & Spector, B. (2011). Experimental evidence for embedded scalar implicatures. *Journal of Semantics*, *28*(3), 359–400.

Chierchia, G. (2004). Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. *Structures and beyond*, *3*, 39–103.

Chierchia, G., Crain, S., Guasti, M., Gualmini, A., & Meroni, L. (2001). The acquisition of disjunction: Evidence for a grammatical view of scalar implicatures. In *Proceedings of the 25th boston university conference on language development* (pp. 157–168).

Chierchia, G., Fox, D., & Spector, B. (to appear). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In P. Portner, C. Maienborn, & K. von Heusinger (Eds.), *An International Handbook of Natural Language Meaning.* Mouton de Gruyter.

Conrad, R., & Hull, A. (1964). Information, acoustic confusion and memory span. *British Journal of Psychology*, *55*(4), 429–432.

De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental psychology*, *54*(2), 128–133.

Engle, R., Tuholski, S., Laughlin, J., & Conway, A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, *128*(3), 309.

Geurts, B. (2006). Take 'five': the meaning and use of a number word. *Non-definiteness and plurality*, 311–329.

Geurts, B., & Der Slik, F. van. (2005). Monotonicity and processing load. *Journal of semantics*, *22*(1), 97.

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

Horn, L. (1972). *On the semantic properties of logical operators in English.* Indiana University Linguistics Club.

Horn, L. (1992). The said and the unsaid. *Ohio State University Working Papers in Linguistics*, *40*, 163–192.

Huang, Y., & Snedeker, J. (2009a). Online interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cognitive psychology*, *58*(3), 376–415.

Huang, Y., & Snedeker, J. (2009b). Semantic meaning and pragmatic interpretation in 5-year-olds: Evidence from real-time spoken language comprehension. *Developmental psychology*, *45*(6), 1723.

Kadmon, N. (1985). The discourse representation of noun phrases with numeral determiners. In *Proceedings of nels* (Vol. 15, pp. 353–422).

Kadmon, N. (2001). *Formal pragmatics: Semantics, pragmatics, presupposition, and focus.* Wiley-Blackwell.

Krifka, M. (1996). Parametrized sum individuals for plural anaphora. *Linguistics and Philosophy*, *19*(6), 555–598.

Levinson, S. (2000). *Presumptive meanings: The theory of generalized conversational implicature.* MIT Press.

Marty, P., & Chemla, E. (2011). *Scalar implicatures: working memory and a comparison with 'only'.* (Ms. LSCP)

Marty, P., Chemla, E., & Spector, B. (2012). Between 3 and 5 *sometimes means* at least 3*: new ways to detect a new ambiguity.* (Ms. LSCP-IJN)

Miyake, A., & Shah, P. (1999). *Models of working memory: Mechanisms of active maintenance and executive control.* Cambridge Univ Pr.

Musolino, J. (2004). The semantics and acquisition of number words: integrating linguistic and developmental perspectives. *Cognition*, *93*(1), 1–41.

Nouwen, R. (2003). *Plural pronominal anaphora in context: dynamic aspects of quantification.* Unpublished doctoral dissertation, LOT.

Noveck, I. (2001). When children are more logical than adults: Experimental investigations of scalar

implicature. *Cognition*, *78*(2), 165–188.
Noveck, I., & Posada, A. (2003). Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language*, *85*(2), 203–210.
Papafragou, A., & Musolino, J. (2003). Scalar implicatures: Experiments at the semantics-pragmatics interface. *Cognition*, *86*(3), 253–282.
Pouscoulous, N., Noveck, I., Politzer, G., & Bastide, A. (2007). A developmental investigation of processing costs in implicature production. *Language acquisition*, *14*(4), 347.
Rooij, R. van, & Schulz, K. (2006). Pragmatic meaning and non-monotonic reasoning: The case of exhaustive interpretation. *Linguistics and Philosophy*, *29*(2), 205–250.
Zondervan, A. (2010). *Scalar implicatures or focus: an experimental approach* (Vol. 249). LOT dissertation series.