Interpreting and Processing Numeral Quantifiers

Paul MARTY
<paul.marty@ehess.fr>

Paris, June 6, 2011
# Contents

1 Theoretical & Empirical Situation 1
   1 Phenomena ............................................. 1
      1.1 The two meanings of bare numerals ................. 1
      1.2 Bare numerals as scalar terms ..................... 3
      1.3 From bare to prepositional numerals ............... 5
   2 Towards a unified account for these phenomena? ............ 6
      2.1 Three options on the table ....................... 6
      2.2 Three questions to be addressed .................. 7
      2.3 Testing the psycholinguistic properties of these phe-
           nomena ............................................. 8

2 Offline Results 9
   1 Experiment 1: ‘between \( n_1 \) and \( n_2 \)’ as a quantifier .... 9
      1.1 Participants and their task ....................... 9
      1.2 Experimental items ................................ 10
      1.3 Procedure ....................................... 12
      1.4 Experimental hypothesis .......................... 12
      1.5 Results ......................................... 12
      1.6 Experiment 1: summary ............................ 15
   2 Experiments 2 & 3 ..................................... 15
      2.1 Experiment 2: ‘between \( n_1 \) and \( n_2 \)’ in a new environment 15
      2.2 Experiment 3: bare numerals as quantifiers ....... 17
   3 A comparison between Experiments 1, 2 & 3 .................. 19
   4 Discussion ........................................... 20

3 Time Course Results 22
   1 Description of Experiment 4 .......................... 23
      1.1 Participants ...................................... 23
      1.2 Experimental items ............................... 23
      1.3 Procedure ...................................... 26
      1.4 Experimental hypothesis .......................... 27
## CONTENTS

2 Results ................................................................. 27  
   2.1 Data treatment ..................................................... 27  
   2.2 Responses to the truth value judgment task ................. 28  
   2.3 Decision and reading times ...................................... 32  
3 Discussion ............................................................... 38  

4 Dual Task Results ................................................. 39  
   1 Description of Experiment 5 ...................................... 40  
      1.1 Participants ..................................................... 40  
      1.2 Material and tasks ............................................ 40  
      1.3 Procedure ....................................................... 43  
      1.4 Experimental hypothesis ..................................... 44  
   2 Results ............................................................... 45  
      2.1 Letter memory task ........................................... 45  
      2.2 Graded truth value judgment task ......................... 46  
   3 Discussion ........................................................... 51  

5 General summary .................................................... 53  

A .......................................................... 55  
   1 Instructions for Experiments 1, 2 & 3 (translated from French) 55  
   2 Target and filler colors ........................................... 56  
   3 Experiment 2: detailed results ................................... 57  
      3.1 Participants ..................................................... 57  
      3.2 Detailed results ................................................ 57  
   4 Experiment 3: experimental items and detailed results ......... 58  
      4.1 Participants ..................................................... 58  
      4.2 Experimental items ............................................ 58  
      4.3 Detailed results ................................................ 60  
   5 Instructions for Experiment 4 (translated from French) ........ 61  
   6 Instructions for Experiment 5 (translated from French) ........ 61
Abstract

The present study is concerned with the interpretation and processing of numeral quantifiers. These phrases have the potential to produce ambiguous readings: a sentence like three Jedi are evil may be interpreted either as at least three Jedi are evil, or as exactly three Jedi are evil. We will start by focusing on a range of comparable ambiguities.

We will first present the ambiguity generated by the use of scalar terms such as ‘some’, for which two readings are also available: a sentence like some Jedi are evil may be interpreted either literally as some, possibly all, Jedi are evil, or through a scalar implicature as some, but not all, Jedi are evil. More surprisingly, we will show that the same type of ambiguity arises with phrases which apparently overtly exclude them: in the right conditions, ‘between n₁ and n₂’ (e.g., between three and five) can be read as ‘at least n₁’. We will explain the underlying theoretical aspects more completely in Chapter 1.

Having established a range of similar ambiguity phenomena on theoretical and introspective grounds, we will investigate and compare these phenomena with quantitative methods. We will explore their cognitive properties with three picture matching studies (Chapter 2), a response time study (Chapter 3), and a dual-task study (Chapter 4).

Overall, the results will show that the phenomena may be fully comparable on introspective and qualitative grounds, but once quantitative and cognitive properties come into the picture, these phenomena show very different profiles. Hence, our results argue for a closer look at the formal implementations explaining each of these ambiguity phenomena. The relevant theories should not only be able to derive the similarity between these phenomena, but also the preference, time course, and memory differences we reveal (see Table 5.1 for a summary of the results).
Chapter 1

Theoretical & Empirical Situation

1 Phenomena

1.1 The two meanings of bare numerals

Bare numerals are known to have different interpretations. Consider the following example:

(1) Three Jedis are evil.

The sentence in (1) is ambiguous between the 'at least' interpretation in (2-a) and the 'exact' interpretation in (2-b):

(2) a. At least three Jedis are evil.
    b. Exactly three Jedis are evil.

Under the interpretation in (2-a), 'three' has a weak, 'lower bounded' meaning equivalent to ‘at least three’ (or ‘three and more’). Alternatively, this word can be interpreted with a strong, 'doubly bound' meaning as in (2-b), under which ‘three’ means ‘exactly three’ (or ‘three and no more’).

One simple way to explain these two interpretations is to introduce two lexical entries for ‘three’. Let us explain how the ‘at least’ interpretation in (2-a) and the ‘exact’ interpretation in (2-b) can be compositionally derived from the following two lexical entries:

(3) a. \[[three] = \lambda P_{(e,t)} \wedge Q_{(e,t)} . \exists x [\# x = 3 \wedge P(x) \wedge Q(x)]\]
    b. \[[three] = \lambda P_{(e,t)} \wedge Q_{(e,t)} . \max \{ n : \exists x [\# x = n \wedge P(x) \wedge Q(x)] \} = 3\]

The lexical entry in (3-a) corresponds to an existential quantification account of numeral determiner denotation (Kadmon 2001): \[[three] , i.e. the denotation of ‘three’, is the property that any plural individual made up of \(P_e\) and \(Q_e\) has iff it contains (at least) three individual members, where \(P\) and \(Q\) are
two predicative noun phrases (NPs) of the semantic type \(\langle e, t \rangle\). The numeral quantifier is first combined with the NP ‘Jedis’ which is taken to denote a property and may be represented by \(\lambda x.\text{jedi}(x)\). The resulting expression is then combined with the property denoted by the second NP ‘evil’ to finish the semantic derivation. The outcome is interpreted by way of composition:

\[
\exists x[\#x = 3 \land \text{jedi}(x) \land \text{evil}(x)]
\]

This says that there is a plural individual of cardinality 3, and this plural individual is made up of Jedis who are evil. This may look like an ‘exact’ reading, but it isn’t. Imagine, for instance, a situation where 6 Jedis are evil; then, the formula in (4) is true in this situation, for if there are 6 evil Jedis, there is necessarily a collection of 3 evil Jedis. This reasoning is based on the fact that ‘Jedis’ and ‘evil’ are both distributive predicates and as such, when applied to plural noun phrases, they distribute over the members of the plurality (Carlson 1977). Therefore, (4) ends up paraphrasing by at least three Jedis are evil, which corresponds to the ‘at least’ interpretation given in (2-a).

Even thought (4) gives a fair representation of the meaning of (2-a), it does not fully capture what such a sentence could normally convey. Typically, an utterance of (1) can also be understood as saying that the cardinality of evil Jedis is exactly three. This alternative ‘exact’ interpretation can be captured by using the lexical entry of ‘three’ in (3-b). Following a similar semantic derivation as the one described above, we obtain the following new outcome for (1):

\[
\max\{n : \exists x[\#x = n \land \text{jedi}(x) \land \text{evil}(x)]\} = 3
\]

This formula is verified iff the maximal number \(n\) of individuals with the properties of being a Jedi and being evil is equal to 3.\(^1\) As a result, the truth-conditions of (5) are such that it is false if the cardinality of the evil Jedis is fewer than three, but also false if more than three Jedis are evil. In sum, it says that there is exactly one plural individual of three evil Jedis, which corresponds to the ‘exact’ interpretation given in (2-b).

In spite of the increasing number of studies on the semantic and pragmatic properties of numerals, there is no firm consensus in the literature on the respective status of these two interpretations: which of them comes first and which is derived? Among the theories accounted for the relations that hold them together, perhaps the most influential view is that number words have

\(^1\)Imagine that our domain is \(\{a, b, c, d\}\), all of which satisfy both \(P\) and \(Q\). Then, \(\max\{n : \exists x[\#x = n \land P(x) \land Q(x)]\} = 3\) is false since the maximal number of individuals having both properties is superior to 3.
a primary ‘at least’ meaning from which a strengthened ‘exact’ meaning can be derived through scalar implicature. Under this standard view, numerals are subject to the same mechanism of strengthening as scalar terms like ‘some’. In the next section, we set out how the different interpretations of numerals are handled within the general framework of scalar implicatures and discuss some observations suggesting that numerals do not behave identically as regular scalar expressions.

1.2 Bare numerals as scalar terms

In the stream stemming from Grice’s seminal work on the pragmatics of human communication (Grice 1975), numerals have been considered as scalar items similar to other quantifiers (e.g., some, many, most). To explain what is at stake, consider the following examples:

(6) a. Three Jedis are evil.
    b. Some Jedis are evil.

Semantically, both numeral and positive quantifiers have been assigned a lower bounded meaning giving rise to an ‘at least’ interpretation (Horn 1972, Grice 1989): three means in fact ‘three or more’ and some means ‘some (possibly all)’. In this analysis, the literal meanings of (6-a) and (6-b) are respectively compatible with a situation in which four and all Jedis are evil. Yet, they are typically understood as implying that not four and not all Jedis are evil. The part of these interpretations involving the falsity of a stronger alternative (i.e., not four and not all) is known as a Scalar Implicature (SI). Thus, the semantic contribution of an utterance of (6-a) has two components:

(7) a. Literal Meaning: Three (possibly more) Jedis are evil.
    b. Scalar Implicature (SI): Not four Jedis are evil.

Put together, (7-a) and (7-b) entail that the number of evil Jedis is exactly three, which results in a doubly bound interpretation.

The traditional view on scalar implicatures (Horn 1972, Grice 1975, Gazdar 1979) is that they are derived from the following pragmatic reasoning. Sentence (6-a) is in competition with the minimally different sentence Four Jedis are evil. This alternative is entertained because of a more basic competition between the lexical items three and four. The four-alternative does not differ much from the uttered sentence but is more informative since it asymmetrically entails the original three-sentence. This asymmetric entailment permits one to view three and four as elements of an implicative scale, \{one, two, three, four, \ldots\}, where terms on the right of the scale are logi-
cally stronger than terms on the left. All things being equal, giving more information is a valuable principle for a cooperative conversation (Grice’s maxim of Quantity), and the fact that the otherwise ‘better’ alternative was not uttered suggests that the speaker was not in a position to deliver the additional information, e.g., because s/he believes that this four-sentence is false. The falsity of the four-alternative corresponds to the SI in (7-b) we were trying to explain.\(^2\)

While this account is widely accepted for the doubly bound interpretations attached to weak positive quantifiers such as ‘some’, it has never really generated a consensus in the case of numeral quantifiers. Actually, there is a great deal of evidence which suggests that numerals behave differently from other scalar terms with respect to their interpretation. Let us illustrate this point with a concrete example.

A characteristic feature of SIs is defeasibility due to structural or contextual constraints (Sadock 1978). For instance, the doubly bound interpretations generated by the assertions in (6-a) and (6-b) can be canceled by additional linguistic material such as follows:

(8) a. Three Jedis are evil. In fact, four of them are.
   b. Some Jedis are evil. In fact, all of them are.

In these examples, the doubly bound interpretation of a weaker scalar term (i.e., three and some) is explicitly denied by the continuation which asserts that the truth of a stronger scale-mate (i.e., four and all) holds. However, when embedded under negation, the doubly bound interpretation of numerals is easier to deny than the SI attached to ‘some’ (Horn 2001); moreover, in the case of ‘some’, the cancellation of the SI under negation requires a special phonological marking (typed in uppercase) to be grammatical:

(9) a. Vador didn’t kill three of the Jedis. He killed four of them.
   b. Vador didn’t kill SOME of the Jedis. He killed ALL of them.

From such observations, it has been argued in recent years that numeral quantifiers have in fact a primary doubly bound meaning from which a lower bounded meaning can be derived. Defenders of this view may differ somewhat on the details of the derivation (Geurts 2006, Breheny 2008), but they agree that the ‘exact’ interpretation of numerals is based on the decoding of their linguistic meaning and, as a result, comes first.

\(^2\)Scalar implicatures can be triggered by numerous syntactic constituents such as connectives (e.g., or, and), modals (e.g., might, must), gradable adjectives (e.g., good, excellent) and so on (see Hirschberg 1991, Horn 2001).
Without a doubt, theoretical inquiries provide a valuable source of insights concerning the difference between numerals and regular scalar expressions. However, we can go further in the investigation of these seemingly different phenomena by comparing their psycholinguistic properties. It is precisely this new step in the study of numeral interpretation we propose to take forward. In addition to this, we will extend our concerns to a third type of ambiguity that have never been described nor investigated in the theoretical and experimental literature. We present this new phenomenon in the next section.

1.3 From bare to prepositional numerals

To the best of our knowledge, the semantics and pragmatics of the prepositional numeral ‘between $n_1$ and $n_2$’ has not yet been studied in detail in the theoretical literature. In this section, we propose an account for the truth conditional contribution of this expression inspired by the semantic analysis of bare numerals given in section 1.1. This proposal will lead us to predict that sentences in which this expression occurs as a quantifier are also ambiguous between an ‘at least’ and an ‘exact’ interpretation. Let us explain the underlying reasoning.

Examining the lexical entries given in (3-a) and (3-b), we observe that the semantic contribution of a bare numeral quantifier has two main components. First, it introduces an existential quantification ($\exists$-), expressing that properties $P$ and $Q$ holds for some (at least one) value of the quantified variable $x$. Second, it provides an information about the cardinality of the set of values satisfying these properties. In the case of bare numerals, this information always consists of a single amount by definition; but, in principle, this cardinal information could be more complex and refer, for instance, to a range of amounts. This is precisely what seems to be at stake in the case of ‘between $n_1$ and $n_2$’.

To explain this point, consider the following sentence obtained from (1) by replacing ‘three’ with ‘between three and five’:

\[(10) \text{Between three and five Jedis are evil.}\]

The semantic contribution of ‘between three and five’ in (10) appears to be very close to the one of ‘three’ in (1). In both cases, the numeral quantifier

---

3The semantics of modified numerals involving spatial prepositions has been discussed in Corver and Zwarts (2006) and Nouwen (2009), but with a focus on the case of under and up to.

4In (3-b), these two components are supplemented by a third one which can be described as a maximality operator, symbolized MAX.
introduces an existential quantification by asserting that there is a plural individual made up of evil Jedis, and provides an information about the cardinality of the members it contains. Crucially, the only point on which they differ is the complexity of the cardinal information they provide: ‘three’ equates the relevant cardinality to a singleton (i.e., \{3\}), ‘between three and five’ equates it to a set of values (i.e., \{3, 4, 5\}).

Thus, we can plausibly assume that bare numerals and ‘between \(n_1\) and \(n_2\)’ are quite similar from the standpoint of their semantic structures. Hence, by analogy with ‘three’, we could expect that two meanings are also available for ‘between three and five’ when this phrase is used as a quantifier:

\[
\begin{align*}
\text{(11)} \quad & \text{a. } \lambda P, Q \langle e, t \rangle \exists x [\# x \in \{3, 4, 5\} \land P(x) \land Q(x)] \\
& \text{b. } \lambda P, Q \langle e, t \rangle \text{Max } \{n : \exists x [\# x = n \land P(x) \land Q(x)]\} \in \{3, 4, 5\}
\end{align*}
\]

By a reasoning analogous to the one sketched for bare numerals (see section 1.1), it results from this assumption that (10) would normally be ambiguous between an ‘at least’ and an ‘exact’ interpretation just like (1). Basically, these two interpretations would lead to different truth values of this sentence in situations where the the cardinality of evil Jedis is superior to five. In such situations, (10) would be false under its ‘exact’ interpretation (12-b) but true under its ‘at least’ interpretation (12-a).

\[
\begin{align*}
\text{(12)} \quad & \text{a. } \exists x [\# x \in \{3, 4, 5\} \land \text{jedi}(x) \land \text{evil}(x)] \\
& \approx \text{At least three Jedis are evil.} \\
& \text{b. } \text{Max } \{n : \exists x [\# x = n \land \text{jedi}(x) \land \text{evil}(x)]\} \in \{3, 4, 5\} \\
& \approx \text{At least three and at most five Jedis are evil.}
\end{align*}
\]

Yet, the ‘at least’ interpretation predicted by our analysis doesn’t seem to be available. However, it could be that such an interpretation exists but is hard to detect. After all, in many cases where it has been claimed in the literature that a certain interpretation does not exist, the interpretation in question has been shown to exist by adequate experimental means (see recently Chemla and Spector 2011). Hence, we hope to design an approach sufficiently powerful to determine whether this ‘at least’ interpretation is or is not available for such sentences — and if so, whether it is available to the same degree as for bare numerals and uncontroversial scalar terms.

2 Towards a unified account for these phenomena?

2.1 Three options on the table

We have three main options to account for the ambiguity phenomena presented above. The first option consists of setting that the items generating
Towards a unified account for these phenomena?

these phenomena are lexically ambiguous. Under this analysis, the relevant items have multiple distinct lexical entries, each of them responsible for a particular interpretation. Arguably, this option is costly and far from satisfactory since it assumes that there are ambiguities throughout the lexicon.

Rather, a more elegant solution is to assume that, when different meanings are available for a same item, these meanings are related to each other. Under this view, the relevant items encode a basic meaning, but appear to be ambiguous because of an optional mechanism which allows to shift from this basic meaning to another one. The function of this mechanism can take two forms with respect to the items under consideration: either such a mechanism is required to derive a doubly bound meaning from a primary lower bounded meaning (second option), or, conversely, to derive a lower bounded meaning from a primary doubly bound meaning (third option).

2.2 Three questions to be addressed

Is it psychologically plausible to adopt the same approach for these three ambiguity phenomena? If so, the same mechanism could account for all of them. As we pointed out, this assumption is at the core of the Scalar Implicature view of numerals (see section 1.2) which argues that the process involved in the derivation of the doubly bound meaning for numerals is similar as the process involved in the derivation of the strengthened meaning for scalar quantifiers such as ‘some’. On the contrary, it could be that these phenomena involved different mechanisms, i.e. mechanisms which differ from a functional point of view.

In the present study, we will contribute to this debate by asking about bare numerals and ‘between $n_1$ and $n_2$’ the same questions that have been asked about ‘some’ in the experimental literature so far:

1) what are the interpretations available for these expressions and to what extent?
2) what is the time course of their derivation?
3) what is the role of the central component of working memory in their derivation?

There are various views on the nature of this mechanism, which may but need not exclude each other. It may be seen as a pragmatic reasoning (see section 1.2), as a type-shifting rule (Geurts 2006, Partee 1987) or as a linguistic operator associated with (possibly covert) lexical elements.
2.3 Testing the psycholinguistic properties of these phenomena

We know how to experimentally tackle the three questions we addressed above. Behavioral studies have investigated for the last decade the why and how of SIs with a great success (for a review, see Noveck and Sperber 2007). Among the various methods used to inquire into their psycholinguistic properties, the truth value judgment task paradigm (TVJT) has proved to be efficient by providing us with a better understanding of their processing time. Using this paradigm, we can for instance ask participants to assess the truth value of a sentence containing a target item such as ‘some’:

(13) Some elephants are mammals.

Crucially, (13) is false with its upper bounding SI (because all elephants are mammals). But this sentence is true under its lower bounded meaning, i.e. without its SI. Hence, in this TVJT example, participants’ responses provide an information as to whether they derived a SI (‘false’ responses) or whether they did not (‘true’ responses). Studies using this paradigm (Noveck and Posada 2003, Bott and Noveck 2004) established that participants took significantly more time to derive the SI than the lower bounded interpretation associated with utterances of (13).6

More recently, De Neys and Schaeken (2007) have suggested that the processing delay observed in the computation of a SI could specifically involve the central component of working memory. Using the TVJT paradigm within a dual task approach, they found that participants derived significantly fewer SIs (approximately 10% less) when their executive resources were tapped, whereas the same cognitive load did not interfere with the comprehension of some non-SI sentences. It confirms that deriving a SI is associated with correlates of effortful processing.

Taking inspiration from these paradigms, we designed three different approaches to study the psycholinguistic properties of the interpretations arising from bare numerals and ‘between $n_1$ and $n_2$’ and compare them with those of the SI attached to ‘some’. We present the results we get from each of them in the following chapters, which are respectively devoted to their offline interpretations (Chapter 2), the time course of their derivations (Chapter 3) and the memory cost they involve (Chapter 4).

---

6This finding, showing that SI derivation comes after an additional processing time, was replicated later by several studies using different methods such as text comprehension task (Katsos et al. 2005, Breheny et al. 2006) or visual-world eye-tracking paradigm (Huang and Suedek 2009).
Chapter 2

Offline Results

In this chapter, we present three studies based on the same experimental design. The main results show that i) a lower bounded or ‘at least’ interpretation is available for sentences in which the expression ‘between $n_1$ and $n_2$’ occurs as a quantifier, and ii) that this interpretation is available to a lower degree than for similar sentences with bare numerals.

1 Experiment 1: ‘between $n_1$ and $n_2$’ as a quantifier

In this experiment, we show that an ‘at least’ reading is available for sentences of the form ‘Between $n_1$ and $n_2$ P are Q’. Such sentences may be judged true when the relevant cardinality is superior to $n_2$.

1.1 Participants and their task

15 native speakers of French ranging in age from 20 to 33 years (mean age 24 years) took part in this experiment (9 women). All of them were native speakers of French and none had any prior exposure to formal linguistics.

Taking inspiration from Chemla and Spector (2011), we designed a graded truth value judgment task for which participants were asked to assess the truth value of a sentence in a situation which was represented graphically. Participants were instructed that sometimes their judgment may not be sharp and that they may thus give their answers along a continuum of answers, by setting with a cursor the length of a red line along a line from ‘No’ to ‘Yes’ (see Figure 2.1). 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. The actual instructions are reported in appendix 1. Responses were coded as a percentage of the line filled in red.
1 Experiment 1: ‘between $n_1$ and $n_2$’ as a quantifier

Figure 2.1: The continuum scale of answers used for the graded truth value judgment task.

1.2 Experimental items

Each item consisted of a sentence and a picture. We describe each of these two components separately below.

Sentences

The sentences of primary interest in the experiment were of the form ‘Between $n$ and $n + 2$ dots are ⟨color⟩’. The numeric term used for $n$ was either 3 or 4, so that $n + 2$ corresponded respectively to 5 or 6. The ⟨color⟩ was an exemplar from a list of ‘target colors’ (see appendix 2). Participants were thus presented Between sentences such as (1):

(1) Between 3 and 5 dots are red.

We are interested in the following potential interpretations of these sentences:

(2) Potential interpretations of (1)
   a. ‘At least’ interpretation: At least 3 dots are red.
   b. ‘Exact’ interpretation: At least 3 and at most 5 dots are red.

Basically, these two interpretations of (1) lead to different truth values in situations where the number of red dots is superior to 5. In such situations, (1) is false under its ‘exact’ interpretation but true under its ‘at least’ interpretation. It is precisely these crucial cases we are interested in.

Participants were also asked to judge parallel sentences with the superlative quantifier At least or At most. Theses sentences were obtained from the Between sentences by replacing the quantifier ‘Between $n$ and $n + 2$’ either with ‘At least $n$’ or with ‘At most $n + 2$’ respectively. A schematic description of the sentence types is provided in Table 2.1. For instance, the following two sentences were obtained from (1):

(3) a. At least 3 dots are red.
    b. At most 5 dots are red.
Experiment 1: ‘between \(n_1\) and \(n_2\)’ as a quantifier

These sentences are not ambiguous: literally, (3-a) is compatible with situations where the minimal number of red dots is 3, and (3-b) with situations where the maximal number of red dots is 5. As a result, in the crucial cases we described above, i.e. when (1) is expected to be ambiguous, (3-a) is true whereas (3-b) is false.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>Between (n) and (n + 2) dots are \langle\text{color}\rangle.</td>
</tr>
<tr>
<td>At least</td>
<td>At least (n) dots are \langle\text{color}\rangle.</td>
</tr>
<tr>
<td>At most</td>
<td>At most (n + 2) dots are \langle\text{color}\rangle.</td>
</tr>
</tbody>
</table>

Table 2.1: Schematic description of the sentence types used in Experiment 1. For a more concrete illustration, you may read \(n\) as 3, \(n + 2\) as 5 and \langle\text{color}\rangle as red.

Pictures

Each picture was composed of four squares containing from 1 to 6 dots. In each square, dots were displayed as on the faces of a dice and depicted either in the target color or in another color from the list of ‘filler colors’ (see appendix 2).

The number of dots in the target color (abbreviated to target dots henceforth) varied from \(n - 3\) to \(n + 5\), where \(n\) still refers to the numeric term used in the sentence. For instance, sentences where \(n = 3\) (e.g., (1), (3-a) or (3-b)) were paired with pictures containing from 0 to 8 target dots. It gave rise to the following three target conditions: Inferior: the cardinality of target dots is inferior to \(n\), Central: the cardinality of target dots is \(n\), \(n + 1\) or \(n + 2\), Superior: the cardinality of target dots is superior to \(n + 2\). The expected truth values of each sentence type as a function of picture condition are given in Table 2.2.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Inferior</th>
<th>Central</th>
<th>Superior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>False</td>
<td>True</td>
<td>‘Exact’ interpretation: False ‘At least’ interpretation: True</td>
</tr>
<tr>
<td>At least</td>
<td>False</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>At most</td>
<td>True</td>
<td>True</td>
<td>False</td>
</tr>
</tbody>
</table>

Table 2.2: Expected truth values of each sentence type as a function of picture conditions.
1.3 Procedure

The instructions were presented first to allow participants to get used to the display by means of two concrete examples (see appendix 1). After that, participants were given three blocks of 90 test items with a short break in between. All target conditions appeared several times in each block. More specifically, for each sentence type, the Inferior condition was presented with 6 different items and Central, as well as Superior, with 12 items (6 in the Group subcondition and 6 in the No-Group subcondition). In each block, the items were presented in random order.

1.4 Experimental hypothesis

The two interpretations investigated for the Between sentences differ as to whether such sentences are or are not appropriate descriptions of situations where the cardinality of target dots is superior to the upper numeral combined with the quantifier (i.e., \(n + 2\)). Hence, if the ‘exact’ interpretation is the only interpretation available for these sentences, then participants’ responses to Between and At most sentences should be very similar in the Superior condition. However, if an ‘at least’ interpretation is also available, we should observe that Between sentences are judged appropriate to a higher degree.

1.5 Results

Mean responses for each sentence type as a function of the cardinality of target dots is given in Fig 2.2. Shapiro-Wilk tests to the pairwise differences reported below revealed that some of them violate the normality assumption required to apply parametric tests. Hence, all statistical analyses of pairwise differences are computed using Wilcoxon rank-sign tests. However, for the
sake of simplicity, statistical analyses of variance were computed using standard parametric analyses (ANOVA). Wilcoxon tests were also computed by item. The item analyses yielded similar results to the subjects analyses.

Figure 2.2: Mean responses (%) for each sentence type as a function of the cardinality of target dots on the picture (n refers to the numeric term used in the sentence). Error bars refer to standard errors.

Control results: *At least* and *At most* sentences.

Participants’ performances to the *At least* and *At most* control sentences are in line with our expectations. *At least* sentences were judged inappropriate ($M = 5\%, SD = 6$) in the Inferior condition and appropriate in the Central ($M = 93\% SD = 8$) and Superior ($M = 92\% SD = 10$) conditions. Mean responses for the *At most* sentences were less clear-cut since they were judged appropriate up to $34\% (SD = 36)$ in the Inferior condition and up to $57\% (SD = 25)$ in the Central condition. However, crucially, they were judged appropriate up to $M = 13\% (SD = 14)$ in the Superior condition.

Overall, these control results show that participants properly performed the task. Most importantly, the mean responses for the *At least* and *At most*
sentences in the critical Superior condition are as we expected. This way, the results for the At most sentences in the Inferior and Central conditions is not a problem for our purposes. Actually, several theories (Geurts and Nouwen 2007, Cummins and Katsos 2010, Nouwen 2010) predict that such sentences, because of the superlative quantifier, are less appropriate when what is ‘under discussion’ (i.e., the cardinality of target dots referred to in the sentence) is related to a definite amount (i.e., the cardinality of target dots on the picture). However, no semantic account makes clear predictions about the asymmetry we found between the At most and At least sentences, so we propose to leave this minor matter aside in the remainder of this study.

Main result: Between sentences are ambiguous

Participants’ responses to the Between sentences revealed that these sentences are ambiguous. Between sentences were judged inappropriate ($M = 5\%$, $SD = 5$) in the Inferior condition and appropriate ($M = 85\%$, $SD = 13$) in the Central condition. In the Superior condition, they were judged appropriate up to $M = 33\%$ ($SD = 32$) with no significant difference between the mean responses for $n + 3$, $n + 4$ and $n + 5$: $F_{1, 15} = .36$, $p = .55$. Crucially, in the Superior condition, participants judged Between sentences appropriate to a significantly lower degree than the At least sentences, but to a significantly higher degree than the At most sentences: $V = 135$, $p < .0001$ and $V = 11$, $p < .01$, respectively.

Hence, in the situations where the cardinality of target dots was superior to $n + 2$, the Between sentences are judged significantly more appropriate than the At most sentences. This finding is consistent with our proposal that the complex numeral quantifier ‘between $n_1$ and $n_2$’ give rise to ambiguities.

Group vs. No-Group: no effect.

Remember that in the Group subcondition, target dots were systematically distributed among the different squares so that one or more groups composed of $n$, $n + 1$ or $n + 2$ elements were easily available, whereas no such group was available in the No-Group subcondition.

Overall, the grouping of target dots did not influence participants’ responses. In the crucial Superior condition, participants’ responses to the Between sentences were similar between the Group and the No-Group subcondition: $M = 31.5\%$ vs. $M = 35.5\%$, $V = 75$, $p = .16$. The same observation stands for the At least and the At most sentences: $V = 31$, $p = .32$ and $V = 70$, $p = .28$, respectively.
1.6 Experiment 1: summary

The main result of this experiment is that participants judged the *Between* sentences appropriate to a significantly higher degree (∼20%) than the *At most* sentences in the Superior condition.

This finding provides support for the existence of an ‘at least’ interpretation of ‘between \(n_1\) and \(n_2\)’ when it is used as a quantifier. These results are explained if we assume that this quantifier is ambiguous between an ‘exact’ and an ‘at least’ interpretation. They are not expected if only the ‘exact’ interpretation exists.

However, as it stands, the present experiment may be not sufficient to ensure that these results reflect a linguistic effect of ambiguity. We noticed, for instance, that participants’ responses to the *At most* sentences were only partially as we expected. This observation could cast doubt on the reliability of the results for this type of sentences in general and, therefore, on the reliability of the comparison between mean responses for *At most* and *Between* sentences which is at the core of our argumentation.

Moreover, one could argue that the comprehension of *Between* sentences is more difficult than that of *At most* sentences: for instance, the *Between* sentences take two numerals contra one for the *At most* sentences. From this assumption, the difference between mean responses for both types of sentences could rather indicate that participants made more errors to the *Between* sentences in the Superior condition.

This line of explanation is unlikely given the clean results we obtained for the other conditions, but we are going to tackle these potential criticisms more directly.

2 Experiments 2 & 3

The task, the procedure and the instructions in Experiments 2 and 3 were the same as in Experiment 1 (see sections 1.1 and 1.3). For reasons of space, we only report here the main results of these experiments. For a detailed description of the experimental designs and results, please refer to the sections 3 and 4 in appendix.

2.1 Experiment 2: ‘between \(n_1\) and \(n_2\)’ in a new environment

In this second experiment, we tested ‘between \(n_1\) and \(n_2\)’ in a different grammatical structure. Experimental items were the same as for Experiment 1 (see section 1.2) except that we replaced *Between* sentence type with a new one, let us call it *Between*\(^*\), of the form ‘The number of ⟨color⟩ dots
is contained between \( n \) and \( n + 2 \). Hence, a *Between* sentence such as (1) corresponds to the following *Between* sentence:

(4) The number of red dots is contained between 3 and 5.

**Main result: *Between* sentences are not ambiguous**

Mean responses for each sentence type as a function of the cardinality of target dots is given in Fig 2.3. In short, we found that participants judged the *Between* sentences appropriate to a lower degree than the *At least* sentences, but to a similar degree than the *At most* sentences: \( V = 119, p < .001 \) and \( V = 67, p = .69 \), respectively.

Hence, in the situations where the cardinality of target dots was superior to \( n + 2 \), the *Between* sentences are judged (in-)appropriate similarly as the *At most* sentences. These results show that participants have no difficulties with the comprehension of the expression ‘between \( n_1 \) and \( n_2 \)’ in general, and more specifically with the derivation of its ‘exact’ meaning.

![Figure 2.3: Mean responses (%) for each sentence type as a function of the cardinality of target dots on the picture.](image)

Figure 2.3: Mean responses (%) for each sentence type as a function of the cardinality of target dots on the picture (\( n \) refers to the numeric term used in the sentence). Error bars refer to standard errors.
Experiment 2: summary

This finding provides further validation of our previous results. Indeed, since we used the same method in both experiments, if the results found for the Between sentences in Experiment 1 was due to a bias in our experimental design, then a similar response pattern should have been observed for the Between* sentences in Experiment 2. Therefore, this additional control ensures that the main result found for Between sentences is not an artifact but does reflect the existence of an ambiguity triggered by an interaction of the quantifier ‘between $n_1$ and $n_2$’ and fine properties of the linguistic environment in which it appears.

Finally, we would like to show that this new finding for ‘between $n_1$ and $n_2$’, although counterintuitive, gets close to what we observe for the interpretation of bare numerals.

2.2 Experiment 3: bare numerals as quantifiers

In this third experiment, the target sentences were of the form ‘$n$ dots are ⟨color⟩’. Participants were thus presented Bare sentences such as (5), investigated for its two potential interpretations (6-a) and (6-b):

(5) 3 dots are red.

(6) Potential interpretations of (5)
   a. ‘At least’ interpretation: 3 dots are red (possibly more).
   b. ‘Exact’ interpretation: 3 dots are red, and no more.

The two interpretations in (6-a) and (6-b) lead to different truth values of (5) in situations where the number of red dots is superior to 3. In such cases, (5) is false under its ‘exact’ interpretation (for then more than 3 dots are red), but true under its ‘at least’ interpretation.

On the pictures, the number of target dots varied from $n - 3$ to $n + 3$, giving rise to the following three target conditions: Inferior: the cardinality of target dots is inferior to $n$, Equal: the cardinality of target dots is equal to $n$, Superior: the cardinality of target dots is superior to $n$.

Main result: Bare sentences are ambiguous

Mean responses for each sentence type as a function of the cardinality of target dots is given in Fig 2.4. In short, we found that participants judged Bare sentences appropriate to a significantly lower degree than the At least sentences and to a significantly higher degree than the At most sentences: $V = 57$, $p < .05$ and $V = 1$, $p < .01$, respectively.
Hence, in the situations where the cardinality of target dots was superior to \( n \), the *Bare* sentences were judged significantly more appropriate (\( \sim 60\% \)) than the *At most* sentences. This finding is consistent with standard semantic accounts of bare numeral quantifiers according to which these items are ambiguous between an ‘exact’ and an ‘at least’ meaning.

**Figure 2.4:** Mean responses (%) for each sentence type as a function of the cardinality of target dots on the picture (\( n \) refers to the numeric term used in the sentence). Error bars refer to standard errors.

**Experiment 3: summary**

Our results confirm that bare numerals are polysemous between a lower bounded and a doubly bound meaning from which an ‘at least’ and an ‘exact’ interpretation can be respectively derived. They also exclude certain readings that have been hypothesized in the literature. For instance, they do not support the availability of an upper bounded or ‘at most’ interpretation: if an ‘at most’ interpretation was also available for numerals, *Bare* sentences would have been judged appropriate to a higher degree in the *Inferior* condition, i.e. in the situations where ‘at most’ \( n \) dots are in the target color.
Our findings thereby disclaim more recent proposals (Horn 1992, Carston 1995, Horn and Ward 2005) according to which the semantic content of numerals is underspecified between an ‘exact’, an ‘at least’ and an ‘at most’ meaning. In particular, the under-specification view predicts a complete symmetry between the ‘at least’ and ‘at most’ interpretations of numerals. Following this prediction, we should have been observed in our experiment that the mean responses for Bare sentences are very similar in the Inferior and in the Superior condition. This prediction is not borne out and the existence of a lexical ‘at most’ meaning is not reflected in our behavioral data.

3 A comparison between Experiments 1, 2 & 3

Mean responses to the At most, Target and At least sentences in the critical Superior condition as a function of experiment is given in Fig 2.5. The Target sentences correspond respectively to Between sentences in Experiment 1, Between* in Experiment 2 and Bare in Experiment 3.

Figure 2.5: Mean responses (%) to the At most, Target and At least sentences in the Superior condition as a function of experiment. Error bars refer to standard errors.
A 3 × 3 ANOVA, with the mean responses given in Fig 2.5 as the dependent measure, revealed a significant interaction of the sentence type (At most vs. Target vs. At least) and the experiment (1 vs. 2 vs. 3): $F(4, 99) = 13.39$, $p < .0001$. These results show that the ‘at least’ interpretation is available to a different degree for the target sentences tested in each experiment. Precisely, they establish that the ‘at least’ interpretation is not available for Between* sentences whereas it is for Between sentences, but to a lower degree than for Bare sentences.

4 Discussion

We designed an experimental approach to investigate the offline interpretation of numeral quantifiers. Instead of asking for absolute judgments of truth or falsity, we offered participants more choices by asking for graded judgments. We aimed to bypass a potential strong preference for a certain reading and get more fine-grained results, which could reveal differences that remained hidden when subjects were given only two options (for similar suggestions, see Meyer and Sauerland 2009, Chemla and Spector 2011). This method allowed us to confirm previous intuitive claims by empirical means and to reveal a previously unnoticed ambiguity.

First, our results confirmed that bare numeral quantifiers (e.g., three) are polysemous between a lower bounded (e.g., at least three) and a doubly bound (e.g., exactly three) meaning (Experiment 3). Importantly, the availability of a third upper bounded meaning (e.g., at most three) for these items was not observed in the present study, contrary to what is assumed by the under-specification view (Carston 1995). Second, they revealed that such an ambiguity also exists for the quantifying expression ‘between $n_1$ and $n_2$’, and further controls ensure that this ambiguity is not an experimental artifact but crucially depends on the linguistic properties of this expression and how it interacts with its environment (Experiments 1 & 2). Finally, we showed that, even though the lower bounded interpretations triggered by both types of items are very comparable, they are not identical from a quantitative point of view: lower bounded interpretations are more readily available for bare numerals than for ‘between $n_1$ and $n_2$’.

We carried out this first round of studies to determine the truth conditional contribution of ‘between $n_1$ and $n_2$’ and compare it with that of bare numerals. Our proposal was that these expressions differ only on the complexity of the cardinal information they provide: bare numerals equates a cardinality to a singleton, whereas ‘between $n_1$ and $n_2$’ equates a cardinality to a set of values. From these premises, we predicted that, if an ‘exact’ and
an ‘at least’ interpretations are available for bare numeral quantifiers, these two interpretations should be available for the quantifier ‘between $n_1$ and $n_2$’ as well. This prediction is borne out by the present results.

From a more theoretical perspective, we showed in Chapter 1 that we can easily account for this new ambiguity by extending the semantic analysis of bare numerals to ‘between $n_1$ and $n_2$’. Supplementary experimental evidence is required to verify the plausibility of having the same approach for these two phenomena. In the case of bare numerals, this ambiguity has been traditionally attributed to their scalar semantic/pragmatic profile and described similarly as those generated by other scalar terms such as ‘some’. We would like to provide more direct tests of this claim and extend our investigations to the case of ‘between $n_1$ and $n_2$’. Scalar implicatures have been experimentally studied with a great success for the last decade. We are going to capitalize on the methods designed to investigate SIs to investigate these other ambiguity phenomena and compare them in a systematic way.
Chapter 3

Time Course Results

How does the ‘exact’ interpretation of numerals occur in real time? According to one common view, numerals (e.g., *three*) encode a lower bounded meaning (e.g., *three or more*) and need to be strengthened (e.g., *three and no more*) via a Scalar Implicature (SI) to give rise to an ‘exact’ or doubly bound interpretation. Experimental studies have previously established that the meaning strengthening of scalar items such as ‘some’ or ‘or’ is not automatic but comes at an extra processing time (Chierchia et al. 2003, Noveck and Posada 2003, Bott and Noveck 2004, Breheny et al. 2006). In the present study, we aim at comparing directly the time course of the doubly bound interpretation of numerals and of the SI attached to the term ‘some’ (in French, ‘certains’).

To that end, we designed a truth value judgment task (TVJT) approach in which participants were asked to read a sentence with either a bare numeral (e.g., *Three dots are red*), ‘between $n_1$ and $n_2$’ (e.g., *Between three and five dots are red*) or ‘some’ (e.g., *Some dots are red*), and then assess its truth value according to a situation represented graphically. In the crucial cases of this TVJT, these sentences would be judged either ‘true’ or ‘false’, depending on which interpretation would be chosen (either lower bounded or doubly bound). Hence, participants’ true/false responses would indicate the interpretation they have chosen, and we could estimate the time course of each interpretation by relating it to their response times.

Results show that the timing of the doubly bound interpretation associated with these three types of sentences differ. First, we found that participants took more time when they derived the SI for the sentences with ‘some’ than when they didn’t (replicating previous findings). Inversely, we found that they took less time for the ‘exact’ than for the ‘at least’ interpretation of the sentences with a bare numeral. Finally, no such difference in response times was observed between the two responses for the sentences with ‘between $n$ and $n + 2$’. We discuss how this new difference between
‘exact’ interpretations of numerals and SIs bares on recent linguistic debates about the basic meaning of numeral quantifiers.

1 Description of Experiment 4

1.1 Participants

33 native speakers of French ranging in age from 19 to 32 years (mean age 23 years) took part in this experiment (24 women). All of them were native speakers of French and none had any prior exposure to formal linguistics.

1.2 Experimental items

Each item consisted of a sentence and a picture. We describe each of these two components separately below.

Sentences

The sentences were of the form Quantifier dots are ⟨color⟩. The sentences of primary interest had as a quantifier either a bare numeral (i.e., n + 1) or the prepositional numeral ‘between n₁ and n₂’ (i.e., Between n and n + 2). The numeric term used for n was either 3 or 4. The ⟨color⟩ was an exemplar from a list of ‘target colors’ (see appendix 2). Participants were thus presented numerically quantified sentences (NQ-sentences henceforth) such as the following:

(1) a. 4 dots are red.
   b. Between 3 and 5 dots are red.

As we established in Chapter 2, these NQ-sentences are ambiguous between a lower bounded and a doubly bound interpretation. These two interpretations lead to different truth values in situations where, for instance, the number of red dots exceeds 5. In such situations, (1-a) and (1-b) are false under a doubly bound interpretation but true under a lower bounded interpretation. We were interested to compare these sentences with variants like (2) that can be interpreted as doubly bound through a SI (SI-sentences henceforth):

(2) Some dots are red.

Crucially, in the situations where all dots are red, (2) is false with its upper-bounding SI (i.e., Not all dots are red) but true under its lower bounded literal meaning (i.e., Some dots are red, possibly all). Thus, we tested these
NQ- and SI-sentences in situations where they are respectively expected to be ambiguous.

Participants were also asked to judge parallel sentences with a comparative quantifier (i.e., ‘More than’, ‘Less than’), a superlative quantifier (i.e., ‘At least’, ‘At most’) or with ‘All’. The More than and Less than sentences were obtained from the Between sentences by replacing ‘Between n and n+2’ either with ‘More than n−1’ or with ‘Less than n+3’ respectively. Similarly, the At least and At most sentences were obtained by replacing ‘Between n and n+2’ either with ‘At least n’ or with ‘At most n+2’. Finally, the All sentences were obtained from the Some sentences by replacing ‘Some’ with ‘All’. These sentences were added to control that participants have no difficulties in interpreting unambiguous sentences. A schematic description of the sentence types is provided in Table 3.1.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bare</td>
<td>( n + 1 ) dots are (color).</td>
</tr>
<tr>
<td>Between</td>
<td>Between ( n ) and ( n + 2 ) dots are (color).</td>
</tr>
<tr>
<td>Some</td>
<td>Some dots are (color).</td>
</tr>
<tr>
<td>All</td>
<td>All dots are (color).</td>
</tr>
<tr>
<td>More than</td>
<td>More than ( n - 1 ) dots are (color).</td>
</tr>
<tr>
<td>Less than</td>
<td>Less than ( n + 3 ) dots are (color).</td>
</tr>
<tr>
<td>At least</td>
<td>At least ( n ) dots are (color).</td>
</tr>
<tr>
<td>At most</td>
<td>At most ( n + 2 ) dots are (color).</td>
</tr>
</tbody>
</table>

Table 3.1: Schematic description of the sentence types used in Experiment 4. For a more concrete illustration, you may read \( n \) as 3 and (color) as red.

Pictures

Each picture was composed of four squares containing from 1 to 6 dots. In each square, dots were displayed as on the faces of a dice and depicted either in the target color or in another color from the list of ‘filler colors’ (see appendix 2).

The crucial experimental factor we manipulated was the number of target dots to obtain the following five target conditions (\( n \) still refers to the numeric term used in the sentence): **Null**: none of the dots are in the target color, **Inferior**: the cardinality of target dots is \( n - 2 \), **Central**: the cardinality of target dots is \( n + 1 \), **Superior**: the cardinality of target dots is \( n + 4 \), **Complete**: all of the dots are in the target color.
For the NQ-sentences, we also manipulated the grouping of target dots when the cardinality of target dots was equal or superior to \( n + 1 \), i.e. in the **Central** and **Superior** conditions. In the **Group** subcondition, the set of target dots were divided so that there was one subset which cardinality was equal to \( n+1 \); however, in the **No-Group** subcondition, it was systematically divided into subsets which cardinality was different from \( n, n + 1 \) and \( n + 2 \). Subsets were then randomly distributed among the different squares of the picture.

Participants were given three blocks of 120 sentence-picture items. The experimental design was set up so that the proportion of expected True/False responses to the truth value judgment task was balanced between and within blocks. The expected truth values of each sentence type as a function of its target picture conditions are given in Tables 3.2-3.4. The number of different items presented in each sentence-picture condition and subconditions is reported in parenthesis. In each block, the test items were presented in random order.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Picture</th>
<th>Inferior</th>
<th>Central</th>
<th>Superior</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Bare</em></td>
<td></td>
<td>False</td>
<td>True</td>
<td>False/True</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(4)</td>
<td>(2 Group, 2 No-Group)</td>
<td>(8 Group, 8 No-Group)</td>
</tr>
<tr>
<td><em>Between</em></td>
<td></td>
<td>False</td>
<td>True</td>
<td>False/True</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(4)</td>
<td>(2 Group, 2 No-Group)</td>
<td>(8 Group, 8 No-Group)</td>
</tr>
</tbody>
</table>

Table 3.2: Expected truth values of *Bare* and *Between* sentences in the Inferior, Central and Superior conditions (number of items in parenthesis).

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Picture</th>
<th>Null</th>
<th>Central</th>
<th>Complete</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Some</em></td>
<td></td>
<td>False</td>
<td>True</td>
<td>False/True</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(4)</td>
<td>(4)</td>
<td>(8)</td>
</tr>
<tr>
<td><em>All</em></td>
<td></td>
<td>False</td>
<td>True</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(4)</td>
<td>(4)</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.3: Expected truth values of *Some* and *All* sentences in the Null, Central and Complete conditions (number of items in parenthesis).
1 Description of Experiment 4

26

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Inferior</th>
<th>Superior</th>
</tr>
</thead>
<tbody>
<tr>
<td>More than</td>
<td>False</td>
<td>True</td>
</tr>
<tr>
<td></td>
<td>(4)</td>
<td>(4 Group, 4 No-Group)</td>
</tr>
<tr>
<td>Less than</td>
<td>True</td>
<td>False</td>
</tr>
<tr>
<td></td>
<td>(4)</td>
<td>(4 Group, 4 No-Group)</td>
</tr>
<tr>
<td>At least</td>
<td>False</td>
<td>True</td>
</tr>
<tr>
<td></td>
<td>(4)</td>
<td>(4 Group, 4 No-Group)</td>
</tr>
<tr>
<td>At most</td>
<td>True</td>
<td>False</td>
</tr>
<tr>
<td></td>
<td>(4)</td>
<td>(4 Group, 4 No-Group)</td>
</tr>
</tbody>
</table>

Table 3.4: Expected truth values of sentences with comparative and superlative quantifiers in the Inferior and Superior conditions (number of items in parenthesis).

1.3 Procedure

At the beginning of each trial, a sentence was presented in the center of the computer screen and remained until participants reported that they had read and understood it by pressing the key ‘p’ on a keyboard. Then, the sentence disappeared and a picture was displayed on the screen. Participants were asked to provide a truth value judgment for the sentence they read in the situation represented on the picture by using two keys on the keyboard (‘q’=no, ‘p’=yes). The general structure of a trial is given in Figure 3.1. Reading time, decision time and response to the truth value judgment task were recorded.

![Figure 3.1: General structure of a trial](image)

Participants were not given specific instructions on how to interpret the sentences but were told that it was essential to perform correctly the task.
that they keep in place their forefingers on response keys (i.e., ‘q’ and ‘p’). Participants were also encouraged to give their responses as quickly as possible, that is to press ‘p’ as soon as they understood the sentence and to press ‘q’ or ‘p’ as soon as they made up their minds. The actual instructions are reported in appendix 5.

Before they began the experiment, participants were invited to start with a short training composed of 10 complete trials. The sentences used for the training were unrelated to the present experimental issue (e.g., There are red dots or There is no red dot) and just aimed to get used participants to the display. Participants were then given three blocks of test items with a short break in between.

1.4 Experimental hypothesis

According to previous experimental studies, the computation of the SI attached to the word ‘some’ comes after an additional processing time (Noveck and Posada 2003, Bott and Noveck 2004, Breheny et al. 2006). We expect first to replicate this finding by showing that the SI (‘false’ responses) for the Some sentences requires more time to be calculated than their lower bounded interpretation (‘true’ responses) in the Complete condition. Then, if the doubly bound or ‘exact’ meaning of numerals is derived through the same process, a claim assumed by scalar implicature theory of numerals, it should require an extra processing effort in the same way; hence, we would expect that participants take longer to answer ‘false’ (doubly bound interpretation) than ‘true’ (lower bounded interpretation) to the Bare sentences in the Superior condition. The same holds for the new Between cases we introduced.

2 Results

2.1 Data treatment

Both reading and decision times were recorded (see Figure 3.1). In order to minimize the effect of outliers, we trimmed the raw data by removing the items with the 2.5% fastest and 2.5% slowest reading times and decision times for any of the two responses (about 10% of the trials were thus removed). All reported means and standard deviations correspond to the remaining raw data. For statistical analyses, these data were log-transformed to reduce positive skewness.
2.2 Responses to the truth value judgment task

SI-sentences

All sentences: SI-control. Proportion of true responses for the All sentences are reported in Figure 3.2. Participants’ performance to the non-SI All sentences were uniformly high, both in the Central ($M = 97\%$, $SD = 4$) and Complete ($M = 99\%$, $SD = 1$) conditions with no reliable difference between the two ($t(32) = -0.87$, $p = .38$). These results show that participants have not encountered difficulties in interpreting unambiguous sentences with a positive quantifier, and more specifically in interpreting (the negation of) the not-all inference which is at the core of the SI derivation for Some sentences.

![Figure 3.2: Proportion of true responses (%) for All sentences as a function of picture condition. Error bars refer to standard errors.](image)

Some sentences are ambiguous. Proportion of true responses for the Some sentences are reported in Figure 3.3. Participants’ performances to the Some sentences in the Null and Central condition were uniformly high, with no difference between the two ($Ms > 97\%$). Crucially, we found that the rate of true answers in the Complete condition ($M = 53\%$, $SD = 7$) was significantly higher than in the Null condition ($M = 1\%$, $SD = 0.7$) and significantly lower than in the Central condition ($M = 97\%$, $SD = 1$):
\[ t(32) = 6.9, \ p < .0001 \] and \[ t(32) = -5.67, \ p < .0001 \], respectively. Hence, results for the Some sentences confirm that these sentences are ambiguous between a lower bounded and a doubly bound interpretation.

![Figure 3.3](image_url)

Figure 3.3: Proportion of true responses (%) for Some sentences as a function of picture condition. Error bars refer to standard errors.

### NQ-sentences

**Comparative and superlative sentences: NQ-control.** Proportion of true responses for the control NQ-sentences are reported in Figure 3.4. Overall, responses to the control NQ-sentences were as expected. Basically, we found that the proportion of true responses was significantly different between the Inferior and Superior condition for each of these types of sentences (all \( ts > 5, \ s \)).

Relying on the expectations given in Table 3.4, we performed further analyses to measure and compare participants’ accuracy to each of these sentences as a function of the semantic profile of the quantifying expressions they involved: a 2 × 2 ANOVA revealed a main effect of the type of numeral modifier (Comparative vs. Superlative), a main effect of the type of monotonicity (Upward vs. Downward)\(^1\) and a significant interaction between the two factors: Modifier type, \( F(1,96) = 23.26, \ p < .0001; \) Mono-

\(^1\)Upward-entailing quantifiers license inferences from subsets to supersets, whereas downward-entailing quantifiers license inferences from supersets to subsets.
tonicity, $F(1, 96) = 52.04, p < .0001$; Modifier type $\times$ Monotonicity interaction: $F(1, 96) = 12.61, p < .001$. The direction of the main effects is that the rate of correct answers is higher for sentences with a comparative (*More than* and *Less than*) than a superlative (*At most* and *At least*) quantifier and higher for sentences with an upward monotone (*More than* and *At least*) than a downward monotone (*Less than* and *At most*) quantifier.

These results suggest that participants’ accuracy to the control NQ-sentences depends on the semantic profile of the quantifying expressions involved and that both the type of numeral modifier and of monotonicity play a central role in that story. These findings replicate previous results from Geurts and van Der Slik (2005) showing that downward-entailing quantifiers built from cardinals are more difficult to process than upward-entailing quantifiers. They are also consistent with recent proposals in the semantic literature arguing that sentences containing superlative quantifiers are more difficult to understand than similar sentences with comparative quantifiers (Geurts and Nouwen 2007, Cummins and Katsos 2010).

Figure 3.4: Proportion of true responses (%) for control NQ-sentences as a function of picture condition. Error bars refer to standard errors.

**Bare and Between sentences are ambiguous.** Proportion of true responses (%) for *Bare* and *Between* sentences are reported in Figure 3.5. Responses confirmed that the *Bare* and *Between* target sentences are ambiguous between a lower bounded and a doubly bound interpretation.
For the Bare sentences, we found that the rate of true answers in the Superior condition ($M = 39\%$, $SD = 41$) was significantly higher than in the Inferior ($M = 1\%$, $SD = 2$) condition and significantly lower than in the Central condition ($M = 98\%$, $SD = 3$): $t(32) = 5.31$, $p < .0001$ and $t(32) = -8.25$, $p < .0001$, respectively. In the same way, we found that the rate of true answers for the Between sentences in the Superior condition ($M = 18\%$, $SD = 28$) was significantly higher than in the Inferior ($M = 2\%$, $SD = 3$) condition and significantly lower than in the Central condition ($M = 94\%$, $SD = 10$): $t(32) = 3.32$, $p < .01$ and $t(32) = -15.05$, $p < .0001$, respectively. Finally, in the crucial Superior condition, we found a significant difference between the rates of true answers for the Bare and Between sentences: $t(32) = -4.23$, $p < .001$. Hence, the lower bounded interpretation is more available for sentences with bare numerals than for sentences with ‘between $n_1$ and $n_2$’. This pattern of results is very similar to the one we observed in our previous experiments.

Figure 3.5: Proportion of true responses (%) for Bare and Between sentences as a function of picture condition. Error bars refer to standard errors.

However, the difference between the rates of true answers for the Between and Less than sentences failed to reach significance in the Superior condition ($t(32) = 1.38$, $p = .17$). Hence, to ensure that the ‘at least’ interpretation of Between sentences is really detected, additional controls with Between* sentences (see Experiment 2) are being tested. Preliminary results go the right way: the rates of true answers for Between* sentences are similarly inferior to 1% in the Inferior and Superior conditions.
Group vs. No-Group: no effect.

In the Central and Superior condition, NQ-sentences were matched with pictures either in the Group or No-Group subcondition. In the Group subcondition, target dots were systematically distributed among the different squares of the picture so that one group composed of exactly \( n + 1 \) elements was easily available, whereas no group of \( n, n + 1 \) or \( n + 2 \) elements was available in the No-Group subcondition.

We ran paired-samples t-tests on differences of the proportion of true responses between the Group and No-Group subcondition for each NQ-sentence type in the Central and Superior condition. None of these pairwise comparisons reach significance (all \( ts < 1.3, \text{ ns} \)). The grouping of target dots does not seem to have influenced participant’s responses to the truth value judgment task.

SI-sentences vs. NQ-sentences.

Remember that the critical condition for the Some sentences corresponds to the Complete condition, while it corresponds to the Superior condition for the Bare and Between sentences. Crucially, we found that the target NQ-sentences give rise to doubly bound interpretations much more readily that Some sentences. A \( 3 \times 2 \) ANOVA revealed a significant interaction of the type of quantifier (Some vs. Bare vs. Between) and the type of condition (Central vs. Critical): \( F(2,160) = 6.93, p < .01 \). These findings extend to adult language previous results from developmental studies (Papafragou and Musolino 2002, Papafragou and Musolino 2003), which showed that 5-year-old children are more successful in drawing doubly bound interpretations triggered by numeral quantifiers than by the scalar item ‘some’.

2.3 Decision and reading times

As in recent studies investigating the time course of numeral quantifier interpretation (Geurts et al. 2010, Cummins and Katsos 2010), decision times (DTs) have proved to be more informative than reading times (RTs). For this reason, reading times will only have a secondary role in the subsequent analyses.

SI-sentences: a delay effect.

Experimental studies have shown that the SI attached to the word ‘some’ requires more time to be derived than its lower bounded interpretation (Bott
and Noveck 2004, Breheny et al. 2006, a.o.). In the present study, we observed a trend in the same direction within the decision times. DTs for *Some* sentences are reported in Figure 3.6.

![Figure 3.6: Decision time (in second) for *Some* sentences as a function of picture condition. DTs in the **Complete** condition are broken down by response/interpretation: *true* responses (lower bounded interpretations) vs. *false* responses (doubly bound interpretations). All other conditions show DTs to correct responses only. Error bars refer to standard errors.](image)

For *Some* sentences in the critical **Complete** condition, participants took more time when they derived a SI (*false* response) than when they did not (*true* response), but this difference failed to reach significance: $M = 1.24\text{s} \text{ vs. } M = 1.16\text{s}$, $t(18) = -.16$, $p = .87$. However, a 2 × 2 ANOVA revealed that the decision time difference between *true* and *false* responses in the critical **Complete** condition is marginally higher than the decision time difference between the **Central** ($M = .96\text{s}$, $SD = .24$) and the **Null** ($M = .84\text{s}$, $SD = .2$) condition: $F(1, 114) = 3.23$, $p = .07$. Importantly, no main effect of response type (*True* vs. *False*) was found: $F(1, 114) = .24$, $p = .62$.

This trend suggests that there is a slight processing cost incurred in the derivation of the SI attached to ‘some’, independent from a potential true/false bias.

---

3Notice that there were only 19 participants out of 33 for who we could directly compare decision times for *false* and *true* responses.
NQ-sentences

Control NQ-sentences  DTs for the control NQ-sentences are reported in Figure 3.7. We found that participant’s reading and decision times for these sentences were both influenced by the type of numeral modifier and its monotonicity property.

![Figure 3.7: Decision time (in second) for correct responses to the control NQ-sentences as a function of picture condition. Error bars refer to standard errors.](image)

The sentences with a superlative quantifier induced significantly longer reading times than those with a comparative quantifier ($t(32) = 3.79$, $p < .001$) and the downward-entailing sentences gave rise to significantly longer reading times than their upward-entailing counterparts ($t(32) = 4.28$, $p < .001$). These results are supported further with the observation of a main effect of the type of numeral modifier (Comparative vs. Superlative) and of monotonicity (Upward vs. Downward) in a $2 \times 2$ ANOVA: Modifier type, $F(1,31) = 29.25$, $p < .0001$; Monotonicity type, $F(1,31) = 28.58$, $p < .0001$. However, no significant interaction between the two effects was found ($F(1,30) = 1.26$, $p = .26$).

Overall, a very comparable pattern of results was observed within decision times. According to a $2 \times 2$ ANOVA, we found a main effect of the type of modifier and of the type of monotonicity for the Inferior condition, with a marginally significant interaction between the two: Modifier type, $F(1,31) =$
Results

26.47, \( p < .0001 \); Monotonicity, \( F(1,31) = 30.75, p < .0001 \); Modifier type × Monotonicity interaction: \( F(1,23) = 3.13, p = .08 \). This interaction reached significance for the Superior condition (\( F(1,30) = 39.48, p < .0001 \)), but no main effect of Modifier type was observed (\( F(1,31) = .63, p = .43 \)).

Taking the analyses of participant’s responses and response times together, the present results establish that the modified numeral quantifiers ‘more than’, ‘less than’, ‘at least’ and ‘at most’ differ from each other with respect to their processing, as measured in this experiment by the rate and the timing of correct/expected interpretations. Our analyses show that the source of this difference lies mainly in two semantic factors which are i) the monotonicity propriety of the quantifying expression, and ii) the type of numeral modifier it involves. The different significant interactions we found between these two factors allow us to rank these quantifiers as a function of their processing complexity as follows: \textit{More than} < \textit{At least} < \textit{Less than} ≤ \textit{At most}.

\textbf{Bare sentences: an opposite delay effect.} Mean decision times for \textit{Bare} sentences are reported in Figure 3.8. Within decision times, we found that participants took marginally less time when they derived a doubly bound interpretation (\textit{false} response) than when they didn’t (\textit{true} response) in the Superior condition: \( M = 1.21s \text{ vs. } M = 1.49s, t(19) = -1.91, p = .07 \).

No such difference between \textit{false} and \textit{true} responses were found between the Inferior (false) and Central (true) conditions: \( M = .93s \text{ vs. } M = .92s, t(32) = .56, p = .57 \).

Furthermore, a \( 2 \times 2 \) ANOVA revealed that the DT difference found between \textit{true} and \textit{false} in the critical Superior condition is significantly bigger than the DT difference between the Inferior and the Central condition: \( F(1,19) = 5, p < .05 \).

As a whole, these results show that, contrary to the SI attached to ‘some’, the doubly bound interpretation of bare numerals is favored in computation over their lower bounded interpretation which comes after an additional processing time.

\footnote{No main effect or significant interaction was found within reading times (all \( Fs < 1, \text{ns} \)).}
Figure 3.8: Decision time (in second) for *Bare* sentences as a function of picture condition. DTs in the **Superior** condition are broken down by response/interpretation. All other conditions show DTs to correct responses only. Error bars refer to standard errors.

**Between sentences: no delay effect.** Mean decision times for *Between* sentences are reported in Figure 3.8. No significant difference within reading and decision times was found between the *true* and *false* responses for *Between* sentences in the critical **Superior** condition: $t(18) = .09, p = .92$ and $t(18) = 1.16, p = .25$, respectively. Given the relatively high rate (82%) of doubly bound interpretation (*false* response) for the *Between* sentences in the **Superior** condition, there were also very few *true* answers to begin with for these analyses.

The DT difference between the **Inferior** and **Central** condition is similar to the DT difference between *true* and *false* response in the **Superior** condition: $F(1, 18) = 2.53, p = .12$.\(^5\) In sum, participants took about the same time for the ‘exact’ than for the ‘at least’ interpretation of the sentences with ‘between $n_1$ and $n_2$’.

\(^5\)The same observations stand within reading times: $F(1, 18) = .06, p = .79$
Figure 3.9: Decision time (in second) for *Between* sentences as a function of picture condition. DTs in the **Superior** condition are broken down by response/interpretation. All other conditions show DTs to correct responses only. Error bars refer to standard errors.

**Group vs. No-Group: no effect.**

We ran paired-samples t-tests on DT difference between the **Group** and **No-Group** subcondition for each type of NQ-sentences in the **Central** and in the **Superior** condition as a function of response type. None of these pairwise comparisons reach significance (all ts < 1.7, ns). Participants took about the same time to answer to the truth value judgment task in both subconditions.

**SI- vs. NQ-sentences**

We conducted a 3 × 2 ANOVA to examine the effects of quantifier type and response type on the DTs for **Some**, **Bare** and **Between** sentences in their respective critical condition. Their interaction failed to reach significance: $F(2, 25) = .88$, $p = .42$. Hence, the DT differences found between the false and true answers for these sentences are too small to be statistically reliable. Arguably, given the opposite delay effects observed for **Some** and **Bare** sentences, the present analysis suffers from the existing imbalance between the number of false and true answers in the critical data.
3 Discussion

We introduced a TVJT approach to probe and compare the timing of the doubly bound interpretation associated with sentences respectively quantified with bare numerals, ‘between $n_1$ and $n_2$’ and ‘some’. The present study makes three contributions regarding this issue.

First, our results replicate previous findings showing that the not-all SI attached to the word ‘some’ is a costly inference which comes after an additional processing time. Second, they reveal that the delay is reversed in the case of bare numerals: we found that participants were faster ($\sim 280$ms) to answer when they derived the doubly bound than the lower bounded interpretation for *Bare* sentences. Finally, no such difference in reading and decision times between these two interpretations was found for sentences with ‘between $n_1$ and $n_2$’.

These findings show that bare numerals, ‘between $n_1$ and $n_2$’ and ‘some’ are different with respect to the time course of their doubly bound reading. In particular, assuming that the time course of an interpretation informs us about the complexity of the computational steps this interpretation involves, the shorter time taken to make the doubly bound interpretation for the *Bare* sentences suggests that the ‘exact’ meaning of numerals is easier to retrieve and process than their ‘at least’ meaning. Does it imply that their ‘exact’ meaning is generated first and their ‘at least’ meaning calculated at later stages in the sentence comprehension?

Our observations from decision times are prima facie consistent with the view that numerals have a primary ‘exact’ meaning, from which an ‘at least’ construal may be secondly derived (Geurts 2006, Breheny 2008). However, this type of behavioral data do not provide sufficiently direct information to determine when these two interpretations arise during incremental processing. Alternatively, the delay we found could also reflect that an additional work is needed for activating their lower bounded interpretation. Explicit verification tasks such as ours might encourage participants to consider less salient interpretations, i.e. interpretations that they would not have spontaneously considered in the normal processing of the sentence (Storto and Tanenhaus 2004). It is logically possible that the ‘at least’ meaning of numerals is computed immediately from the input, but that it takes time for the resulting lower bounded interpretation to become available and to influence participant’s behavior (Huang and Smeeker 2009, Grodner et al. 2010).

In the last experiment, we propose to investigate further the psychological differences between these ambiguity phenomena by comparing the component of the cognitive cost associated with the derivation of their different interpretations.
Chapter 4

Dual Task Results

Some researchers have argued that the doubly bound or ‘exact’ meaning of numerals is attributable to a Scalar Implicature (SI) which arises from the same process that provides the doubly bound for scalar items such as ‘some’ (Krifka 1999, Winter 2001, a.o.). Capitalizing on recent findings showing that the derivation of a SI draws on working memory resources (De Neys and Schaeken 2007, Marty and Chemla 2011), we designed a dual task approach to compare the computation of the ‘exact’ meaning of numerals and the derivation of the SI attached to ‘some’ (in French, ‘certains’).

Participants were asked to perform a graded truth value judgment task as in Experiment 1 on quantified sentences with either bare numerals (e.g., *Three dots are red*), ‘between n and n + 2’ (e.g., *Between three and five dots are red*) or ‘some’ (e.g., *Some dots are red*), while they simultaneously tried to remember a sequence of letters (see Figure 4.1). 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.

Results show that the working memory resources are not similarly recruited in the derivation of the doubly bound interpretations of these three types of sentences. Whereas tapping participant’s memory resources does not affect their interpretation of ‘between’ sentences, it interferes with their interpretation of the sentences with ‘some’ and bare numerals, but in an opposite way: under high cognitive load, participants made fewer SIs for the ‘some’ sentences (replicating previous findings), whereas they made more ‘exact’ interpretations for the sentences with bare numerals. These new findings establish that, although the ambiguity phenomena generated by these phrases may seem nearly identical, they are very different from the standpoint of their cognitive proprieties.
1 Description of Experiment 5

1.1 Participants

The participants were 26 native speakers of French, aged between 19-34 years (mean age 23 years; 13 females). All of them were native speakers of French and none had any prior exposure to formal linguistics.

1.2 Material and tasks

Letter Memory Task

The dual task was a classical storage task of letters. Input sequences consisted of either 2 or 4 letters in an uppercase font that were presented in the center of the screen one at a time for 1 s each, with a 500 ms blank screen between each letter. The nine letters used were B, F, H, J, L, M, Q, R, and X, chosen to be phonologically distinct. Participants memorized the sequence and were asked afterwards to reproduce the letters in backward order by using a keyboard.

It was established that the cognitive effort required for encoding such a sequence depends on several factors such as, for instance, the phonological similarity (Conrad and Hull 1964, Baddeley 1966) or the number of the elements that it contains (Baddeley et al. 1975). 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 that we used respectively in the low load and in the high load trials. The working memory resources should be minimally burdened by the storage of 2-letter sequences but more tapped by the storage of 4-letter sequences.

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 the dual-task coordination is widely considered as one of its main functions (Baddeley 1992, Miyake and Shah 1999, Engle et al. 1999). The rationale for this dual task procedure is that participants should not be able to appeal to their working memory for the graded truth value judgment task in conditions where the cognitive burden is high, i.e. in conditions where working memory is needed to perform the memory dual task. Hence processes that specifically draw on these resources should be blocked.

To manipulate the load factor within subjects, 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.
Graded Truth Value Judgment Task

Each item consisted of a sentence and a picture. For each sentence-picture item, participants were asked to assess the truth value of the sentence according to the situation represented on the picture. Participants gave their answers by setting with a cursor the length of a red line along a fixed line from ‘No’ to ‘Yes’. Responses were coded as a percentage of the line filled in red.

Sentences

The sentences were of the form *Quantifier dots are ⟨color⟩*. The sentences of primary interest had as a quantifier either a bare numeral \((n + 1)\) or the complex expression ‘between \(n_1\) and \(n_2\)’ \((\text{Between } n \text{ and } n + 2)\). The numeric term used for \(n\) was either 3 or 4. The \(⟨\text{color}⟩\) was an exemplar from a list of ‘target colors’ (see appendix 2). Participants were thus presented NQ-sentences such as the following:

\[
\begin{align*}
(1) & \quad \text{a. 4 dots are red.} \\
& \quad \text{b. Between 3 and 5 dots are red.}
\end{align*}
\]

As we established in Chapters 2 and 3, these NQ-sentences are ambiguous between a lower bounded and a doubly bound interpretation. We compared these sentences with variants like (2) that can be interpreted either literally as lower bounded or through a SI as doubly bound:

\[
(2) \quad \text{Some dots are red.}
\]

Participants were also asked to judge parallel sentences with the quantifier *More than, Less than or All*. *More than* and *Less than* sentences were obtained from the *Between* sentences by replacing ‘Between \(n\) and \(n + 2\)’ either with ‘More than \(n - 1\)’ or with ‘Less than \(n + 3\)’ respectively. *All* sentences were obtained from the *Some* sentences simply by replacing ‘Some’ with ‘All’. These sentences were added to control that the cognitive load did not interfere with the comprehension of unambiguous sentences. A schematic description of the sentence types is provided in Table 4.1.

Pictures

Each picture was composed of four squares containing from 1 to 6 dots. In each square, dots were displayed as on the faces of a dice and depicted either in the target color or in another color from the list of ‘filler colors’ (see appendix 2).
Table 4.1: Schematic description of the sentence types used in experiment 4. For a more concrete illustration, you may read $n$ as 3 and (color) as red.

The crucial experimental factor we manipulated was the number of target dots to obtain the following five target conditions ($n$ still refers to the numeric term used in the sentence): **Null:** none of the dots are in the target color, **Inferior:** the cardinality of target dots is $n - 2$, **Central:** the cardinality of target dots is $n + 1$, **Superior:** the cardinality of target dots is $n + 4$, **Complete:** all of the dots are in the target color.

For the NQ-sentences, we also manipulated the grouping of target dots when the cardinality of target dots was equal or superior to $n + 1$, i.e. in the **Central** and **Superior** conditions. In the **Group** subcondition, the set of target dots were divided so that there was one subset which cardinality was equal to $n + 1$; in contrast, in the **No-Group** subcondition, it was systematically divided into subsets which cardinality was different from $n$, $n + 1$ and $n + 2$. Subsets were then randomly distributed among the different squares of the picture.

Participants were given two blocks of 96 sentence-picture items. The experimental design was set up so that the proportion of expected True/False responses to the graded truth value judgment task was balanced between as well as within blocks. The expected truth values of each sentence type as a function of its target picture conditions are given in Tables 4.2-4.4. The number of different items presented in each sentence-picture condition and subcondition is reported in parenthesis. In each block, the test items were presented in random order.
1 Description of Experiment 5

Table 4.2: Expected truth values of *Bare* and *Between* sentences in the Inferior, Central and Superior conditions (number of items in parenthesis).

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Picture</th>
<th>Inferior</th>
<th>Central</th>
<th>Superior</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Bare</em></td>
<td></td>
<td>False</td>
<td>True</td>
<td>False/True</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(8)</td>
<td>(4 Group, 4 No-Group)</td>
<td>(4 Group, 4 No-Group)</td>
</tr>
<tr>
<td><em>Between</em></td>
<td></td>
<td>False</td>
<td>True</td>
<td>False/True</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(8)</td>
<td>(4 Group, 4 No-Group)</td>
<td>(4 Group, 4 No-Group)</td>
</tr>
</tbody>
</table>

Table 4.3: Expected truth values of *Some* and *All* sentences in the Null, Central and Complete conditions (number of items in parenthesis).

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Picture</th>
<th>Null</th>
<th>Central</th>
<th>Complete</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Some</em></td>
<td></td>
<td>False</td>
<td>True</td>
<td>False/True</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(4)</td>
<td>(4)</td>
<td>(8)</td>
</tr>
<tr>
<td><em>All</em></td>
<td></td>
<td>False</td>
<td></td>
<td>True</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(4)</td>
<td></td>
<td>(4)</td>
</tr>
</tbody>
</table>

Table 4.4: Expected truth values of *More than* and *Less than* sentences in the Inferior and Superior conditions (number of items in parenthesis).

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Picture</th>
<th>Inferior</th>
<th>Superior</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>More than</em></td>
<td></td>
<td>False</td>
<td>True</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(4)</td>
<td>(4 Group, 4 No-Group)</td>
</tr>
<tr>
<td><em>Less than</em></td>
<td></td>
<td>True</td>
<td>False</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(4)</td>
<td>(4 Group, 4 No-Group)</td>
</tr>
</tbody>
</table>

1.3 Procedure

Each trial started with the presentation of the sequence of letters to be memorized. Then, a sentence-picture item was displayed on the computer screen and remained until the participant provided a truth value judgment. Next, participants had to reproduce the sequence of letters in backward order. At the end of each trial, they received feedback on the quality of their sequence reproduction. When the sequence was not correctly reproduced, participants
were reminded that they had to do their best to remember and completely reproduce it (see Figure 4.1).

Participants were instructed that it was essential 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. The actual instructions are reported in appendix 6. Before they began the experiment, they were invited to start with a short training composed of 4 complete trials (2 with 2-letter and 2 with 4-letter sequences). The sentences used for the training were unrelated to the present experimental issue and just aimed to get used participants to the display (e.g., There are red dots). Participants were then given two blocks of test items with a short break in between.

1.4 Experimental hypothesis

Previous experimental studies (De Neys and Schaeken 2007, Marty and Chemla 2011) showed that the executive working memory resources are involved in the derivation of the SI attached to the scalar item ‘some’. Precisely,
it was found that participants derive fewer SIs under high cognitive load. If the doubly bound or ‘exact’ meaning of numerals is derived through the same process, a claim assumed by scalar implicature theory of numerals, it should involve executive memory resources in the same way. Hence, if participants judge the *Some* sentences more appropriate (fewer SIs) in the *Complete* condition when they are under high cognitive load, we would expect that they judge *Bare* and *Between* sentences more appropriate (fewer doubly bound interpretations) in the *Superior* condition as well.

2 Results

2.1 Letter memory task

Overall, the dual task was properly performed. The mean number of correctly reproduced letters was 1.9 out of 2 ($SD = .07$) for the simple 2-letter sequences and 3.8 out of 4 ($SD = .12$) for the complex 4-letter sequences. The reproduction of a letter sequence was scored as correct if the participant reproduced the sequence correctly in its entirety. According to a *t*-test for dependent samples, the difference between the rates of correct reproduction for simple ($M = 96, SD = 4$) and complex ($M = 92, SD = 6$) sequences is significant: $t(25) = -4.17, p < .001$. It confirms that the 4-letter sequences were more demanding than the 2-letter sequences.

Moreover, the correlation between the rate of correct reproduction and the mean responses to *Bare* sentences in the *Superior* condition is small and not significant in low as well as in high load trials ($r = -.27, p = .16$ and $r = .11, p = .57$, respectively). The same holds for the correlation between the rate of correct reproduction and the mean responses to *Some* sentences in the *Complete* condition ($r = -.07, p = .69$ and $r = -.09, p = .63$, respectively). Hence, there does not seem to be a trade-off between letter recall performances and graded truth value judgments.

In all subsequent analyses, trials where participants did not accurately reproduce the complete sequence were removed (about 6% of the trials). In order to minimise the effect of outliers arising through lapses in concentration, we also considered the items with the 2.5% fastest and 2.5% slowest response times as outliers and removed them from the analyses (about 5.5% of the responses). According to a monofactorial analysis of variance (ANOVA), the mean number of removed trials did not significantly differ from one sentence type to another in low as well as in high load trials ($F(5, 125) = 1.53, p = .18$ and $F(5, 125) = .58, p = .71$, respectively).
2.2 Graded truth value judgment task

SI-sentences

Mean responses for All and Some sentences are respectively reported in Fig 4.2 and Fig 4.3.

Figure 4.2: Mean responses (%) for All sentences as a function of target picture conditions in low load and high load trials. Error bars refer to standard errors.

**All sentences: SI-control.** Participants’ performance to the non-SI All sentences was as expected. All sentences were judged appropriate up to 1% ($SD = 4$) in the Central condition and up to $M = 99\%$ ($SD = 5$) in the Complete condition with no difference between the high load and the low load trials: $t(25) = -1.21, p = .23$ and $t(25) = -1.24, p = .22$, respectively. No significant interaction between the type of condition (Central vs. Total) and the type of load (Low vs. High) was found: $F(1, 25) = .12, p = .72$.

These results show that the cognitive load did not interfere with the comprehension of unambiguous sentences with a positive quantifier, and more specifically with the comprehension of the not-all inference which is at the core of the SI derivation for Some sentences.
Some sentences: a replication of the memory effect. Responses to the Some sentences showed that these sentences give rise to upper-bounding SIs. In the Complete condition, Some sentences were judged appropriate ($M = 85\%$, $SD = 15$) to a significantly higher degree than in the Null condition ($M = 0.5\%$, $SD = 1$), but to a significantly lower degree than in the Central condition ($M = 98\%$, $SD = 5$): $t(25) = -32.45$, $p < .0001$ and $t(25) = 5.18$, $p < .0001$, respectively. These results are fully explained iff we assume that Some sentences are ambiguous between a lower bounded and a doubly bound interpretation.

Crucially, in the Complete condition, we found that Some sentences were judged significantly more appropriate in the high load than in the low load trials: $M = 89\%$ vs. $M = 81\%$, $t(25) = 2.68$, $p < .01$. No such difference in mean responses between low load and high load trials was observed in the Null and the Central condition. This memory effect is specific to the crucial Complete condition: a $3 \times 2$ ANOVA revealed a significant interaction between the type of condition (Null vs. Central vs. Complete) and the type of load (Low vs. High) on the mean responses to Some sentences: $F(2, 50) = 5.52$, $p < .01$. 

Figure 4.3: Mean responses (%) for Some sentences as a function of target picture conditions in low load and high load trials. Error bars refer to standard errors.
Hence, the upper-bounding SI that arises from the Some sentences is less available under high cognitive load. This finding replicates previous results (De Neys and Schaeken 2007, Marty and Chemla 2011) and confirms that some of the cognitive processes involved in the computation of SIs draw on executive working memory resources.

NQ-sentences

Mean responses for More than and Less than sentences are reported in Fig 4.4 and mean responses for Bare and Between sentences in Fig 4.5.

![Figure 4.4: Mean responses (%) for More than and Less than sentences as a function of target picture conditions in low load and high load trials. Error bars refer to standard errors.](image)

**More than and Less than sentences: NQ-controls.** Overall, participants’ performances to these sentences are as expected. More than sentences were judged appropriate up to 2% ($SD = 7.6$) in the Inferior condition and up to $M = 97\%$ ($SD = 5.4$) in the Superior condition. This difference is significant: $t(25) = -73.8$, $p < .0001$. Conversely, Less than sentences were judged significantly more appropriate in the Inferior condition than in the Superior condition: $M = 93\%$ vs. $M = 21\%$, $t(25) = 11.88$, $p < .0001$. 
Furthermore, results showed that the concurrent memorization of the complex 4-letter sequence did not interfere with the comprehension of these NQ-sentences. In the Inferior condition, the mean responses for More than and Less than sentences were about the same in low load and high load trials: $t(25) = -1.23$, $p = .22$ and $t(25) = -0.63$, $p = .53$, respectively. The same observations hold in the Superior condition: $t(25) = -1.68$, $p = .1$ and $t(25) = -.32$, $p = .7$, respectively. Finally, no significant interaction between the type of condition (Inferior vs. Superior) and the type of load (Low vs. High) was found for both types of sentences (all $F$s < 0.2, ns).

In sum, these results show that the cognitive load did not interfere with the comprehension of unambiguous sentences with numeral quantifiers.

Figure 4.5: Mean responses (in %) for Bare and Between sentences as a function of picture conditions in low load and high load trials. Error bars refer to standard errors.

**Bare sentences: an opposite memory effect.** Results confirmed that Bare sentences are ambiguous between a lower bounded and a doubly bound interpretation. In the Superior condition, Bare sentences were judged appropriate ($M = 46\%$, $SD = 38$) to a significantly higher degree than in the Inferior condition ($M = 1\%$, $SD = 4$), but to a significantly lower degree than in the Central condition ($M = 98\%$, $SD = 3$): $t(25) = -6.25$, $p < .0001$ and $t(25) = 7.04$, $p < .0001$, respectively.
Crucially, in the critical Superior condition, we found that Bare sentences were judged significantly more appropriate in the low load than in the high load trials: \( M = 54\% \) vs. \( M = 38\% \), \( t(25) = 4.33, p < .001 \). Furthermore, a 3 × 2 ANOVA revealed that the mean response difference found between low load and high load trials in the Superior condition is significantly bigger than the mean response differences between low load and high load trials in the Inferior and the Central condition: \( F(2, 50) = 16.61, p < .0001 \).

The concurrent memorization of the 4-letter sequences interferes with the comprehension of the Bare sentences, inducing a significant decrease of lower bounded interpretations (\(~15\%)\). Hence, participants made more ‘exact’ interpretations of bare numerals under high cognitive load.

**Between sentences: no effect of the dual task.** Results for Between sentences provide further evidence to support that these sentences are ambiguous between a lower bounded and a doubly bound interpretation. In the Superior condition, Between sentences were judged appropriate \( (M = 26\%, \ SD = 15) \) to a significantly higher degree than in the Inferior condition \( (M = 2.5\%, \ SD = 1) \), but to a significantly lower degree than in the Central condition \( (M = 94\%, \ SD = 5) \): \( t(25) = 3.51, p < .001 \) and \( t(25) = -8.81, p < .0001 \), respectively.\(^1\) We found that the difference between mean responses for Bare and Between sentences in the crucial Superior condition is significant: \( t(25) = 4.94, p < .0001 \). This finding is consistent with our previous results (Experiments 1 – 3) showing that an ‘at least’ reading is available for the quantifier ‘between \( n_1 \) and \( n_2 \)’ but to a lower degree than for bare numerals.

We conducted paired-samples t-tests to compare participant’s difference mean responses between high load and low load trials in each target picture condition. None of these tests reach significance (all \( ts < 1.8 \), ns). Nor significant interaction between the type of load (Low vs. High) and the type of picture condition (Inferior vs. Central vs. Superior) was found: \( F(2, 50) = 0.24, \ p = .78 \). The concurrent memorization of the 4-letter sequences does not seem to have influenced participants’ judgments to the Between sentences.

\(^1\)In contrast with Experiment 4 (see footnote 2), we also found that the Between sentences were judged appropriate to a significantly higher degree than the Less than sentences in the Superior condition, both in low load and high load trials: \( t(25) = 5.78, p < .0001 \) and \( t(25) = 3.2, p < .01 \), respectively.
3 Discussion

Group vs. No-Group: no effect.

Remember that, in the Central and Superior conditions, NQ-sentences were matched with pictures either in the Group or No-Group subcondition. In the Group subcondition, target dots were systematically distributed among the different squares of the picture so that one group composed of exactly \(n+1\) elements was easily available, whereas no group of \(n\), \(n+1\) or \(n+2\) elements was available in the No-Group subcondition.

We ran paired-samples t-tests on differences of mean responses between the Group and No-Group subcondition for each NQ-sentence type in the Central and the Superior condition. None of these pairwise comparisons reach significance in the low load as well as high load trials (all \(ts < 1.2, \text{ns}\)). Finally, no significant interaction between the type of picture subcondition (Group vs. No-Group) and the type of load (Low vs. High) was found both in the Central and Superior condition (all \(Fs < 1, \text{ns}\)).

NQ-sentences vs. SI-sentences

Remember that the critical condition for Bare and Between sentences corresponds to the Superior condition, while it corresponds to the Complete condition for Some sentences. We found that the effect of the dual task was different for each of these types of sentences in the critical conditions: according to a \(3 \times 2\) ANOVA, there is a significant interaction of the type of quantifier (Bare vs. Between vs. Some) and the type of load (Low vs. High): \(F(2,50) = 13.35, \ p < .0001\). These results show that the doubly bound interpretations triggered by ‘some’, bare numerals and ‘between \(n_1\) and \(n_2\)’ differ from each other in terms of memory demands.

3 Discussion

We designed a dual task approach to modulate, during interpretation of ambiguous SI- and NQ-sentences, the availability of processes involving the central component of working memory. Our goal was to know whether these processes are or are not similarly involved in these comparable ambiguity phenomena. The results make three contributions regarding this issue.

First and foremost, our results replicate previous findings showing that the SI attached to ‘some’ comes at a memory cost. Then, they show that such a memory cost is absent in the derivation of comparable interpretations associated with sentences quantified with either bare numerals or ‘between \(n_1\) and \(n_2\)’. Precisely, we found that tapping participants’ executive resources induces a significant decrease of SIs for sentences with ‘some’, whereas it
generates a significant increase of ‘exact’ interpretations for sentences with bare numerals and has no effect on sentences with ‘between $n_1$ and $n_2$’. What can we learn from these results?

Working memory resources are known to constrain on the ability to maintain multiple interpretations during the resolution of linguistic ambiguities (Miyake et al. 1994). If the dual task effect was simply to strengthen the salience of the logically stronger readings, i.e. the lower bounded readings, we should have notably observed in the high load trials an increase of the mean responses for the Bare sentences in the Superior condition, as we found for the Some sentences. Instead, we found that participants made significantly fewer upper bounded interpretations for this type of sentences under high cognitive load.

Hence, assuming that participants tried to be as efficient as possible within available resources, the dual task effect was to inhibit the costlier computations in terms of memory resources, i.e. the computation of the doubly bound interpretation for ‘some’ and the computation of the lower bounded interpretation for bare numerals. These new findings go against the view that the strengthened meaning of numeral quantifiers and regular scalar items would be obtained through similar processes, and reveal fundamental differences between the cognitive properties of these ambiguity phenomena.
Chapter 5

General summary

The ambiguity phenomena generated by bare numerals have been standardly assumed to be similar to those generated by regular scalar terms such as ‘some’, for which two readings are also available. We used offline results to quantify this ambiguity of numerals. We also showed for the first time that, under the right conditions, the more complex phrase ‘between \( n_1 \) and \( n_2 \)’ gives rise to comparable ambiguities. In particular, it gives rise to ‘at least’ readings, despite the fact that the upper bound \( (n_2) \) is specified explicitly. Hence, we carried out further investigations to explore and compare the fine-grained psycholinguistic properties of these seemingly comparable phenomena. A general summary of our results is given in 5.1.

<table>
<thead>
<tr>
<th>Properties</th>
<th>Phrase</th>
<th>bare numerals</th>
<th>‘between ( n_1 ) and ( n_2 )’</th>
<th>‘some’</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detection of DB*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp 1 &amp; 3</td>
<td>bare numerals</td>
<td>22%</td>
<td>67%</td>
<td>—</td>
</tr>
<tr>
<td>Exp 4</td>
<td></td>
<td>61%</td>
<td>82%</td>
<td>47%</td>
</tr>
<tr>
<td>Exp 5 (low load trials)</td>
<td></td>
<td>46%</td>
<td>75%</td>
<td>19%</td>
</tr>
<tr>
<td>Time course (Exp 4)</td>
<td>LB &gt; DB</td>
<td>LB ≈ DB</td>
<td>LB &lt; DB</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(slower)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Memory cost (Exp 5)</td>
<td>LB &gt; DB</td>
<td>LB ≈ DB</td>
<td>LB &lt; DB</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(‘harder’)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.1: General summary of the results. The labels LB and DB respectively refer to Lower Bound and Doubly Bound reading.

*Rejection rate of the relevant sentence in a situation in which the DB reading is false and the LB reading is true.

The similarity between the three phenomena seemed to call for a unified account of these ambiguities. However, our quantitative and psycholinguistic results show that these three types of phenomena have very different profiles...
regarding the salience (Experiments 1−5), time course (Experiment 4) and memory cost (Experiment 5) of the different interpretations they involve.

First, the computation of the doubly bound interpretation (SI) attached to ‘some’ is costlier than the computation of its lower bounded counterpart, both in terms of processing time and working memory resources. Conversely, in the case of bare numerals, these correlates of effortful processing are associated with the computation of their lower bounded interpretation. Finally, no timing or memory differences was found between the two interpretations of ‘between \(n_1\) and \(n_2\)’. As a result, these quantifying expressions may be ranked as a function of the cognitive effort required for computing their strengthened meaning as follows: \textit{bare numerals} \(<\textit{‘between }n_1\textit{ and }n_2\textit{’} \(<\textit{‘some’}}

These new findings provide psycholinguistic evidence which should shed new light on formal implementations of each of these ambiguity phenomena. The psycholinguistic data extend the comparison allowed by standard offline introspective data. They go against the view that the doubly bound reading of numerals and scalar items is derived through the same mechanism (see the theoretical options discussed in Chapter 1, section 2.1). Our results suggest that the doubly bound reading attached to ‘some’ is derived from a literal lower bounded meaning, as assumed by the the traditional view on Scalar Implicatures (see Chapter 1, section 1.2); on the contrary, the lower bounded reading of bare and modified numerals would be derived from a primary doubly bound meaning, as suggested by more recent proposals (Geurts 2006, Breheny 2008).
Appendix A

1 Instructions for Experiments 1, 2 & 3 (translated from French)

Thank you for your participation to this experiment. You are going to see sentences and situations in which dots are depicted in various colors. Your task is to tell whether the sentence is true or false in this situation. For instance, in the example below, the sentence is true:

![Diagram showing at least 3 dots are red.](image)

Important: this is not a math test! In fact, in many cases that you will see, the sentence will not be clearly true or clearly false, but it will describe more or less appropriately the situation, will be more or less natural in this situation. Hence, there is no good answer, and we are mainly interested in your intuition. For this reason, you can give more finegrained judgments than a simple ‘yes’ or ‘no’: you will give your answers by setting the length of a red line along a line from ‘No’ to ‘Yes’. The more the sentence seems appropriate, the more you will extend the line to the right with the mouse, close to ‘Yes’. This will certainly be the case for examples like the one above.
2 Target and filler colors

In Table A.1, we present the list of target and filler colors we used in experiments 1, 2 and 3 (actual items were in French). For each experimental item, the color referred to in the sentence was randomly chosen from the list of target colors. Then, the filler color used in the picture was pseudo-randomly chosen from the list of filler color minus the selected target color.
3 Experiment 2: detailed results

In Experiment 2, we tested ‘between $n_1$ and $n_2$’ in a different grammatical structure and show that, in these new environment, this expression is not ambiguous and can only give rise to ‘exact’ interpretations.

3.1 Participants

11 native speakers of French ranging in age from 18 to 26 years (mean age 20 years) took part in this experiment (10 women). All of them were native speakers of French and none had any prior exposure to formal linguistics.

3.2 Detailed results

Participant’s responses

Overall, participants’ performances to the At least and At most control sentences were similar to the ones previously observed in Experiment 1. At least sentences were judged inappropriate ($M = 3\%$, $SD = 7$) in the Inferior condition and appropriate in the Central ($M = 94\%$, $SD = 10$) and Superior ($M = 92\%$, $SD = 13$) conditions. At most sentences were judged appropriate up to $37\%$ ($SD = 38$) in the Inferior condition and up to $64\%$ ($SD = 40$) in the Central condition. However, as expected, they were judged appropriate up to $M = 14\%$, ($SD = 34$) in the critical Superior condition.

Participants’ responses to the Between* sentences showed that these sentences were not ambiguous. Between* sentences were judged inappropriate ($M = 3\%$, $SD = 9$) in the Inferior condition and appropriate ($M = 86\%$, $SD = 27$) in the Central condition. In the Superior condition, they were judged appropriate up to $M = 10\%$ ($SD = 28$) with
a significant difference between the mean responses for \( n + 3 \), \( n + 4 \) and \( n + 5 \): \( F(2, 27) = 4.64, p < 0.05 \). Pairwise post-hoc comparisons of the means (Tuckey’s HSD) showed that the mean for \( n + 5 \) was significantly lower than the means for \( n + 3 \) and \( n + 4 \) (\( ts > 2.5, s \)). Crucially, in the Superior condition, participants judged the Between* sentences appropriate to a lower degree than the At least sentences, but to a similar degree than the At most sentences: \( V = 119, p < .001 \) and \( V = 67, p = .69 \), respectively.

**Group vs. No-Group: no effect.**

We found that the grouping of target dots did not influence participants’ responses to our graded judgment task. In the Superior condition, participants’ responses to the Between* sentences were similar between the Group and No-Group subcondition: \( M = 11\% \) vs. \( M = 9\% \), \( V = 44, p = .34 \). The same observation stands for the At least and the At most sentences: \( V = 23, p = .15 \) and \( V = 37, p = .1 \), respectively.

## 4 Experiment 3: experimental items and detailed results

In Experiment 3, we confirm by empirical means previous intuitive claims that an ‘at least’ interpretation is available for sentences of the form ‘\( n P_s \) are \( Q_s \)’, where \( n \) represents a non-modified number word, also called bare numeral.

### 4.1 Participants

11 native speakers of French ranging in age from 18 to 30 years (mean age 22 years) took part in this experiment (8 women). All of them were native speakers of French and none had any prior exposure to formal linguistics.

### 4.2 Experimental items

**Sentences**

Target sentences were of the form ‘\( n \) dots are \( \langle color \rangle \)’. The numeric term used for \( n \) was 3, 4, 5 or 6. At least and At most sentences were respectively obtained by appending the superlative quantifier ‘At least’ or ‘At most’ to the Bare sentences. A schematic description of the sentence types is provided in Table A.2.
4 Experiment 3: experimental items and detailed results

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bare</td>
<td>(n) dots are (\langle\text{color}\rangle).</td>
</tr>
<tr>
<td>At least</td>
<td>At least (n) dots are (\langle\text{color}\rangle).</td>
</tr>
<tr>
<td>At most</td>
<td>At most (n) dots are (\langle\text{color}\rangle).</td>
</tr>
</tbody>
</table>

Table A.2: Schematic description of the sentence types used in experiment 3. For a more concrete illustration, you may read \(n\) as 3 and \(\langle\text{color}\rangle\) as red.

Pictures

As in Experiments 1 and 2, we manipulated the number of target dots. It varied from \(n - 3\) to \(n + 3\), giving rise to the following three target conditions: Inferior: the cardinality of target dots is inferior to \(n\), Equal: the cardinality of target dots is equal to \(n\), Superior: the cardinality of target dots is superior to \(n\). The expected truth values of each sentence type as a function of picture conditions are encapsulated in Table A.3.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Picture</th>
<th>Inferior</th>
<th>Equal</th>
<th>Superior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bare</td>
<td>False</td>
<td>True</td>
<td>‘Exact’ interpretation: False</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>‘At least’ interpretation: True</td>
<td></td>
</tr>
<tr>
<td>At least</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td></td>
</tr>
<tr>
<td>At most</td>
<td>True</td>
<td>True</td>
<td>False</td>
<td></td>
</tr>
</tbody>
</table>

Table A.3: Expected truth values of each sentence type as a function of picture conditions.

We also manipulated the grouping of target dots in situations where their cardinality was superior or equal to \(n\), i.e. in the Equal and the Superior condition. In the Group subcondition, the set of target dots was divided into subsets so that there was one subset composed of exactly \(n\) elements; conversely, in the No-Group subcondition, it was systematically divided into subsets which cardinality was different from \(n\). Subsets were then randomly distributed among the different squares of the picture. As a result, in the Group subcondition, a group of exactly \(n\) target dots was easily available on the picture, whereas no such group was available in the No-Group subcondition.

In all, participants were given two blocks of 132 test items with a short break in between. All target conditions appeared several times in each block. More specifically, for each sentence type, the Inferior condition was presented with 12 different items, Equal with 8 (4 in the Group and 4 in the No-Group subconditions).
No-Group subcondition), and Superior with 24 items (12 in the Group and 12 in the No-Group subcondition). In each block, the items were presented in random order.

4.3 Detailed results

Participant’s responses

Participants’ performances to the At least and At most control sentences were similar to the ones previously observed in Experiment 1 and 2. At least sentences were judged inappropriate (\(M = 3\%\), \(SD = 5\)) in the Inferior condition and appropriate in the Equal (\(M = 89\%\) \(SD = 14\)) and Superior (\(M = 91\%\) \(SD = 14\)) conditions. At most sentences were judged appropriate up to 40\% (\(SD = 30\)) in the Inferior condition. However, as expected, they were judged appropriate up to 93\% (\(SD = 7\)) in the Equal condition and up to \(M = 18\%\), \(SD = 26\) in the Superior condition.

Participants’ responses to the Bare sentences showed that these sentences were ambiguous. They were judged inappropriate (\(M = 4\%\), \(SD = 6\)) in the Inferior condition and appropriate (\(M = 99\%\), \(SD = 1\)) in the Equal condition. In the Superior condition, these sentences were judged appropriate up to \(M = 78\%\) (\(SD = 19\)) with no significant difference between the mean responses for \(n + 1\), \(n + 2\) and \(n + 3\): \(F(2, 30) = .01, p = .98\). Crucially, in the Superior condition, participants judged Bare sentences appropriate to a significantly lower degree than the At least sentences and to a significantly higher degree than the At most sentences: \(V = 57, p < .05\) and \(V = 1, p < .01\), respectively.

Group vs. No-Group: no effect.

We ran Wilcoxon tests to the pairwise differences of mean responses between the Group and No-Group subcondition for each sentence type in each picture condition. No significant difference was found (all \(ts < 1.4, ns\)). As in Experiments 1 and 2, there does not seem that participants’ responses were influenced by the grouping of target dots.
5 Instructions for Experiment 4 (translated from French)

Thank you for your participation to this experiment. You are going to see sentences from French. Each time a sentence is displayed on the screen, you have to read it. As soon as you have understood what the sentence means, press the key ‘YES’ (key ‘p’ on the keyboard). You will then see a picture in which dots are depicted in various colors. Your task is to tell whether the sentence is true or false in the situation represented on the picture:

- If, according to you, the sentence describes appropriately the situation, press the key ‘YES’ (key ‘p’ on the keyboard).
- On the other hand, if the sentence does not describe appropriately the situation, press the key ‘NO’ (key ‘q’ on the keyboard).

To carry out successfully this task, you must put and keep in place your forefinger on response keys (‘Yes’ and ‘No’) in order to be able to give your responses quickly: press ‘YES’ as soon as you have understood the sentence and give your response ‘No’/‘Yes’ as soon as you have made your mind.

Sometimes the sentences look like each other; this does not matter, always answer following your intuition for each example, independently of your previous answers. This experiment is not long, but it is a bit repetitive and you have to stay focused. Even if you need to be attentive, do not spend too much time on each question, follow your intuition.

If you have any questions before beginning, please feel free to ask them us. The first sentences will help you to get used to the display and to the task. Press ‘Space’ when you are ready.

6 Instructions for Experiment 5 (translated from French)

Thank you for your participation to this experiment. Letters of the alphabet are going to appear, one at a time, in the center of the screen. Your task is to memorize these letters so as to reproduce them later in the reverse of the order in which they originally appear. For instance, if the letters A, B, C, D are successively displayed on the screen, you have to reproduce the sequence of letters D C B A. To do so, use the standard keyboard in front of you and press the keys corresponding to the letters you want to reproduce. Your letters will be automatically displayed on the screen as in the example below:
If you think that you have not correctly reproduce the sequence of letters, you can delete your current answer to restart by pressing the left arrow key or clicking on the button ‘Delete’. However, if you think that your reproduction is correct, press the right arrow key or click on the button ‘Confirm’ to confirm your answer and continue. A message will briefly appear informing you on the quality of your reproduction.

After the presentation of the letters and before their reproduction, you will see sentences associated to situations in which dots are depicted in various colors. Here is an example of what you will see:

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

You might hesitate, but follow your feeling and use the flexibility of the red bar at best to represent your intuition about the correspondence between the sentence and the situation. You will get used quickly and intuitively to this bar.

If you have any questions before starting, please feel free to ask. When you are ready, press ‘Space’ bar to start.
Bibliography


De Neys, W. and W. Schaeken (2007). When people are more logical un-


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