The likelihood of upper-bound construals among numeral modifiers[1]

Yaron McNabb — Utrecht University
Stavroula Alexandropoulou — Utrecht University
Dominique Blok — Utrecht University
Sofia Bimpikou — Utrecht University
Rick Nouwen — Utrecht University

Abstract.

In this paper we show that modified numerals differ with respect to the nature of the bounds they express. We examine the numeral modifiers *less/fewer than*, *at most* and *up to* in a series of experiments in English and Greek, and investigate to what extent these modifiers impose an upper bound. Our results indicate that the upper-bound construal that *up to* gives rise to is cancellable in contrast with the uncancellability of the upper-bound construal that *at most* and *less/fewer than* give rise to. This finding is compatible with an analysis that treats the upper bound of *at most* and *less/fewer than* as part of their semantic content and the upper bound of *up to* as a pragmatic inference. In addition, we discuss the effect of the scalar distance between possible alternatives and the modified numeral on the likelihood and strength of the upper-bound construal.

Keywords: Numeral modifiers, scalar implicature, experimental semantics and pragmatics.

1. Introduction

Numeral modifiers provide a fruitful case study of several theoretically-relevant semantic and pragmatic phenomena, specifically ignorance inferences, free choice inferences, scalar implicatures, and interaction with granularity (Geurts and Nouwen, 2007; Büring, 2008; Cummins and Katsos, 2010; Nouwen, 2010; Schwarz, 2011; Cummins et al., 2012; Schwarz, 2013; Kennedy, 2013, 2015; Cohen and Krifka, 2014)

The focus of our experimental investigation is the diversity of scalar inferences, particularly upper-bound construals, among different modified numerals. Blok (2015) claims that modified numerals differ in how their upper-bound inferences are derived: the upper bound of *at most* and *less/fewer than* is part of the semantic content whereas the upper bound of *up to* is derived pragmatically. In order to test these intuitions, we developed a task that utilizes a modified Likert scale in order to obtain gradient judgements on the strength of the upper bound construal, on the assumption that consistent responses to an upper-bound reading point to it being semantic, while varied or gradient

[1]For discussion of material related to this article, we thank the audience in Sinn und Bedeutung 20 in Tübingen, as well as audiences at the Experimental Approached to Semantics (ExAS) workshop at ESSLLI 2015 in Barcelona, XPRAG 2015 in Chicago, and the MXPRAG 2015 workshop in Berlin, where previous versions of this study were presented. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement no. 313502.

responses point to it being pragmatic.

In Section 2, we motivate the experimental investigation of the three numeral modifiers. We begin in Section 2.1 with the initial observation in Geurts and Nouwen (2007) that numeral modifiers divide into two classes based on whether they give rise to ignorance inferences: *up to* and *at most* belong to the same class as they both give rise to ignorance inferences, while *less/fewer than* belongs to a different class, as it doesn't. Despite this class distinction, however, modifiers from neither class give rise to scalar implicature, in contrast with bare numerals, as discussed in Section 2.2, unless granularity comes into play, as discussed in Section 2.3. But if the upper bound of bare, but not modified, numerals, is derived via scalar implicature, how is it derived in the modified numeral *up to* when granularity isn't at stake? A look at the monotonicity of the modifiers reveals that *at most* and *less/fewer than* are downward-monotone, while *up to* isn't, suggesting that the upper bound of the first two is entailed while the upper bound of the latter is derived via a pragmatic inference (Section 2.4). On the assumption that pragmatic inferences are cancellable and lead to variable responses, as discussed in Section 2.5, we test the nature of the upper bound in the three numeral modifiers in Greek (Section 3) and English (Section 4). In Section 5, we discuss additional factors that may contribute to the degree of upper-bound cancellability.

2. Inferences numeral modifiers give rise to

2.1. Ignorance inferences

Geurts and Nouwen (2007) show that superlative modifiers like *at most* give rise to ignorance inferences. Nouwen (2010) argues that the absence or occurrence of such inferences is the hallmark of an essential distinction between two classes of scalar quantifiers. CLASS A numeral modifiers like *more than, less/fewer than, under* and *over* don't give rise to ignorance inferences. This is why explicitly expressing knowledge of an exact number, e.g., how much memory the speaker's laptop has in (1a), can be followed by a numeral modified by this class of modifiers. In contrast, CLASS B modifiers like *at least, at most, minimally, maximally* and *up to* do lead to ignorance inferences, and so following up an utterance with explicitly-communicated speaker certainty with a numeral modifier of this class is infelicitous, as in (1b).

(1) a. **Class A**

I know exactly how much memory my laptop has, and it's $\left\{\begin{array}{c}\text{more than}\\ \text{less than}\\ \text{under}\\ \text{over}\end{array}\right\}$ 4GB.

b. **Class B**

#I know exactly how much memory my laptop has, and it's $\left\{\begin{array}{c}\text{at least}\\ \text{at most}\\ \text{minimally}\\ \text{maximally}\\ \text{up to}\end{array}\right\}$ 4GB.

As will be shown in the following sections, while ignorance inferences constitute a point of variation among numeral modifiers, all of these expressions behave similarly with respect to scalar implicature. Specifically, none of them give rise to scalar implicatures when they are not embedded under certain operators or when fine granularity is involved.

2.2. Scalar implicature

The classic analysis of number words is that they have a one-sided, lower-bound only meaning, and that the exact, upper- as well as lower-bound meaning is derived by scalar implicature. The listener reasons that by asserting *three* in (2), the speaker doesn't know if greater numbers, e.g., four, hold. Strengthening this implicature, such that the speaker knows that greater numbers don't hold, leads to the implicature in (2), following '$\rightsquigarrow$,' which stands for 'is used to implicate' (Horn, 1972; Gazdar, 1979; Levinson, 2000; Sauerland, 2004; Geurts, 2010).[2]

(2) John has three children.
$\rightsquigarrow$ John doesn't have four children.

Krifka (1999) and Fox and Hackl (2006) observe that when the number word combines with a numeral modifier like *more than*, the upper-bound inference doesn't hold, as illustrated by (3a), where '$\not\rightsquigarrow$' stands for 'doesn't implicate'. Additional downward-entailing numeral modifiers like *at most* and *fewer than* don't give rise to scalar implicature, either, as shown in (3b) and (3c) below.

(3) a. John has more than three children.
$\not\rightsquigarrow$ John doesn't have more than four children.
b. At most ten people died in the crash.
$\not\rightsquigarrow$ It's not the case that at most nine people died in the crash.
c. Fewer than ten people died in the crash.
$\not\rightsquigarrow$ Fewer than nine people died in the crash.

[2]That said, amassing evidence and arguments suggest a two-sided analysis of numbers (Horn, 1992; Musolino, 2004; Geurts, 2006; Huang et al., 2013; Kennedy, 2013).

The prevailing family of accounts for the lack of scalar implicature says that sentences with modified numerals are evaluated against alternatives with other modified numerals substituted for the modified numeral in the sentence, e.g., *fewer than* and *exactly*, would be substituted for *at most*.[3] The alternatives for sentences like (2) or (3) are symmetric; that is, they can't be simultaneously false while the assertions in (2) or (3) are true, and so the implicatures don't arise.

However, comparative and superlative modifiers give rise to scalar implicatures in the scope of certain operators, such as certain modals, universal nominal quantifiers, distributive conjunctions, and the quantifier *more than half of the NPs* (Mayr, 2013). Similarly, as will be discussed in the following section, both comparative and superlative modifiers trigger scalar implicatures when granularity is taken into consideration.

2.3. Granularity and scalar distance

Cummins et al. (2012) observe that under certain circumstances, some upper-bound inferences are available from utterances with comparative and superlative modifiers. For example, if Ahmed were born in Cairo, the statement in (4) would be semantically true but intuitively misleading, as *more than 1000* and *at least a 1000* seem to convey a quantity that is less than the actual population of Cairo (about 10 million).

(4) Ahmed's birthplace has $\left\{\begin{matrix}\text{more than}\\ \text{at least a}\end{matrix}\right\}$ 1000 inhabitants.

Cummins et al. (2012) find that the range of numbers communicated by utterances with modified numerals depends on the granularity of the numeral. When a speaker uses a non-round, precise number like 93, she signals to the speaker that this number should be interpreted as 'exactly 93.' That is, a number like 93 has a fine-granularity interpretation. When a speaker, however, uses a round number like 110, it's likely interpreted as a range of values that includes 110; that is, it receives a medium-granularity interpretation. And finally, when a speaker uses a round number like 100, the number is likely interpreted as a greater range of potentially-communicated values than 110, thus receiving a coarse-granularity interpretation (Krifka, 2007). This strategic communication of the approximate interpretation of numbers indeed affects the interpretation of modified numerals. In an experimental investigation, Cummins et al. found that the interpreted upper bound for *more than 100*, which represents coarse granularity, was 149, the upper bound for *more than*

[3]The various accounts differ in the details of how alternatives are computed, but the resulting inferences are the same. Schwarz (2011, 2013) assumes the Horn set {*at least, exactly, at most*} of scalar modifiers in addition to the Horn set of numerals. Mayr (2013) assumes the Horn sets {*at least, at most*} and {*less/fewer than, more than*} for the relevant numeral modifier and the numeral scale. Kennedy (2013) proposes that in the alternatives the numeral is kept constant while *at least n* is substituted by *more than n* (and *at most n* is substituted by *less than n*) and the bare numeral, for which he assumes a two-sided semantics.

110, which represents medium granularity, was 127.5, and the interpreted upper bound for *more than 93*, which represents fine granularity, was 100. Results for the superlative modifier *at least* revealed similar patterns.

In the last two sections, we saw that certain Class A and Class B numeral modifiers behave similarly with respect to scalar implicatures. In the following section we'll focus on Class B numeral modifiers and examine the differences among them with respect to monotonicity and bounds.

2.4. Monotonicity and boundedness

Schwarz et al. (2012) observe that negative polarity items are licensed in the scope of *at most*, but not in the scope of *up to*, concluding that the former, but not that latter, should be analyzed as downward monotone. Based on this and other observations, they conclude that there must be a fundamental semantic difference between these two modifiers. Building on this observation, Blok (2015) argues that this crucial difference is (in part at least) due to the fact that the upper bound expressed by *up to* is implicated rather than entailed. This accounts for the fact that the upper bound set by *at most 23 people* in (5a) cannot be cancelled, whereas the upper bound set by *up to 23 people* in (5b) is cancellable.

(5) a. At most 23 people came to the party, # if not 24.
b. Up to 23 people came to the party, if not 24.

Blok further proposes that while *at most n* denies the existence of occurrences of values higher than *n*, *up to n* asserts the existence of values between some implicit lower bound and *n*. Higher values are only excluded by implicature. Based on a survey of 15 different languages, Blok concludes that the contrast between the counterparts of *at most* and *up to* is a crosslinguistically-consistent contrast.

2.5. Degree of cancellability and strength of implicature

Generalized conversational implicatures and specifically scalar implicatures have been treated as a categorical phenomenon. This assumption dates back to (Grice, 1975), who says that "the use of a certain form of words in an utterance would normally (in the ABSENCE of special circumstances) carry such-and-such an implicature or type of implicature." There is, however, evidence from the processing literature that scalar implicatures are not computed by default (Breheny et al., 2006). Similarly, in Bott and Noveck (2004), a pragmatically-enriched interpretation (i.e., 'some but not all') of the sentence in (6) would contradict the fact that all elephants are mammals, but a logical

interpretation (i.e., 'some and possibly all') would not. However, even participants who were trained to interpret *some* as 'some but not all' accepted the sentence in (6) in 60% of the cases, in contrast with a 85% acceptance rate in the logical condition.

(6) Some elephants are mammals.

What these results show is that the presence of an implicature leads to a certain level of variability in responses. In what follows, we will use this characteristic to test the nature of the upper bound expressed by *up to*. If the upper bound interpretation of *up to* is derived via scalar implicature, then we expect a variable response pattern for (7a) given (8), suggesting that (7a) and (8) can be compatible just in case the upper bound implicature is cancelled. We expect (7b) and (7c), in contrast, to be always incompatible with (8), as the upper bound is part of the semantic content and therefore can't be cancelled (see Blok, 2015 for *at most* and Fox and Hackl, 2006; Nouwen, 2010 for *less/fewer than*).

(7) a. Interns in advertisement companies get **up to** 980 dollars per month.
b. Interns in advertisement companies get **less than** 980 dollars per month.
c. Interns in advertisement companies get **at most** 980 dollars per month.

(8) The interns in some of them are paid **985** dollars per month.

3. Experiment 1: Greek

3.1. Research questions

In Experiment 1, we set to answer two research questions: (i) is the upper bound of *up to* cancellable, in support of an implicature-based account; and (ii) if so, to what extent? In order to investigate the degree of upper-bound cancellability of *up to*, we compare it with *at most* and *less/fewer than*, on the assumption that the upper-bound inference in the latter two is part of the semantic content and is therefore consistent.

3.2. Methods

Experiment 1 was conducted in Greek. The Greek directional numeral modifier equivalent to *up to* is the preposition *mehri*. *To poli* (lit. 'the much') is the Greek counterpart of *at most*, and the adjective *lighoteros/-i/-o apo* (lit. 'fewer.MASC/FEM/NEUT than') and the adverb *lighotero apo* correspond to *fewer than* and *less than*, respectively.

In the experimental task, we asked participants to rate how coherent sentence continuations like the one in (9) were on a modified Likert scale of –3 to 3, where –3 was a "very bad" continuation and 3 was a "very good" continuation. After the instructions and before the actual questionnaire, participants were presented with two practice items, one coherent discourse and one contradictory discourse, which had the form of the filler items (see (10) and (11) below). This served to familiarize participants with the procedure. The first sentence in each pair of sentences was adapted from naturally-occurring tokens gleaned from the Hellenic National Corpus (`http://hnc.ilsp.gr/en/default.asp`).

In the target items, the first sentence was always a general claim, while the second sentence was a more specific claim, highlighting one of the cases the first, general claim was referring to. The first sentence included a number n modified by a numeral modifier; n was almost exclusively a non-round number. The second sentence had a number m, which was either slightly smaller than n in the first sentence ($m < n$; the 'under' condition) or slightly greater than n ($m > n$; the 'over' condition).

(9) Example target item:

I askumeni stis dhyafimistikes eteries pernun { lighotero apo / to poli / mehri } n dholaria ton mina; i askumeni se arketes apo aftes plirononde m dholaria ton mina.

'Interns in advertisement companies get { less than / at most / up to } n dollars per month; the interns in some of them are paid m dollars per month.'

Is the underlined sentence a good continuation of the first sentence?

–3	–2	–1	0	1	2	3
very bad						very good

Thus, in our 3×2 design, conditions differed with respect to the choice of modifier (*lighotero/lighoteros/-i/-o apo* 'less/fewer than', *to poli* 'at most', *mehri* 'up to') in the first sentence and with respect to whether the number in the underlined continuation was smaller or greater than the number in the first sentence (discrepancy factor: 'under' and 'over' levels). The comparative numeral modifier and 'under' were the reference levels of these two factors, respectively.

All target items (N=12) were rotated through six lists in a latin square design, so that each participant only saw one condition per item. Fourteen filler items including seven coherent discourses and seven contradictory discourses were added into the mix, too (see translated example items below), making a total of 26 stimuli.

(10) Example of coherent filler item:
Several countries have more than one official language; for example, Belgium has three official languages: Dutch, French and German.

(11) Example of contradictory filler item:
The Panhellenic examinations started at the end of May; specifically, the examination of the first subject took place on the 10th of June.

218 participants filled in an online questionnaire created on `www.surveymonkey.com`. Data from 67 participants were discarded, because they did not fill in the entire questionnaire or they were not native speakers of Greek. Eight additional participants were excluded, as they gave scores greater than –1 to the contradictory fillers and smaller than 1 to the coherent fillers. Data from the remaining 143 participants (98 female participants, 2 didn't specify gender; mean age: 32.8; age range: 19–67) were used for the statistical analyses reported on here.

3.3. Predictions

Schwarz et al. (2012) argue that the upper bound of both *at most* and *up to* is entailed and would therefore predict no difference in the degree of upper-bound construals between these two numeral modifiers. Blok (2015), on the other hand, argues that *at most*, but not *up to*, entails an upper bound and thus predicts a difference in the degree of upper-bound inferences between the two numeral modifiers. If we find that participants are more likely to approve of an 'over' item when the modifier is *mehri* 'up to' than when it is *to poli* 'at most', this will support Blok's (2015) analysis. On the assumption that *less/fewer than* imposes a semantic upper bound as well (Hackl, 2000; Nouwen, 2010), we would expect participants to make the same distinction between *lighotero apo* 'less than' and *mehri* 'up to'. In the 'under' condition, since the values in the continuation are entailed by the first sentence with the modified numeral, we would expect no difference among the modified numeral conditions.

3.4. Results

The data obtained by 143 participants, summarized in Figure 1, were analyzed with mixed-effects ordered probit regression models using the `ordinal R` package (Christensen, 2013). The full model with modifier, discrepancy, and their interaction as fixed effects, with random intercepts and slopes for modifier, discrepancy, and their interaction for subjects, and with random intercepts and slopes for items was found to be the best fit for our data ($LRstatistic = 22.68$, $df = 4$, $p < .001$). This model showed a marginally significant difference between the modifiers *lighotero apo* 'less than' and *mehri* 'up to' ($\beta = -.244$, $SE = .127$, $p = .055$), a significant difference between the 'under'

and 'over' conditions ($\beta = -1.532$, $SE = .171$, $p < .0001$), and a significant interaction for *mehri* 'up to' and 'over' ($\beta = .472$, $SE = .139$, $p < .001$).

We investigated the modifier effect for each of the two discrepancy conditions with two additional models. Both models had modifier as the only fixed effect. The analysis for the 'over' condition revealed that items with *mehri* 'up to' received significantly higher coherence rates than items with *lighotero apo* 'less than' ($\beta = .226$, $SE = .092$, $p < .05$) or *to poli* 'at most' ($\beta = .306$, $SE = .093$, $p < .001$). No difference was found between *lighotero apo* 'less than' and *to poli* 'at most' ($\beta = -.08$, $SE = .093$, $p <= .389$). In the analysis for the 'under' condition, items with *mehri* 'up to' were found to be borderline significantly different from *lighotero apo* 'less than', with the former scoring lower ($\beta = -.245$, $SE = .146$, $p = .093$), and significantly lower than the *to poli* 'at most' items ($\beta = -.291$, $SE = .111$, $p < .01$). Again, there was no difference in the scores for *lighotero apo* 'less than' and *to poli* 'at most' ($\beta = .046$, $SE = .157$, $p = .769$).

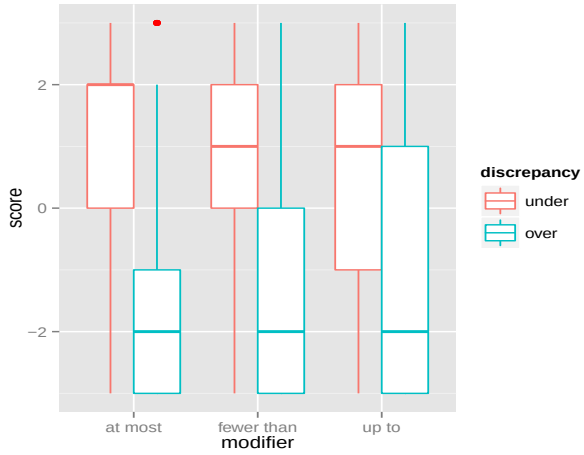
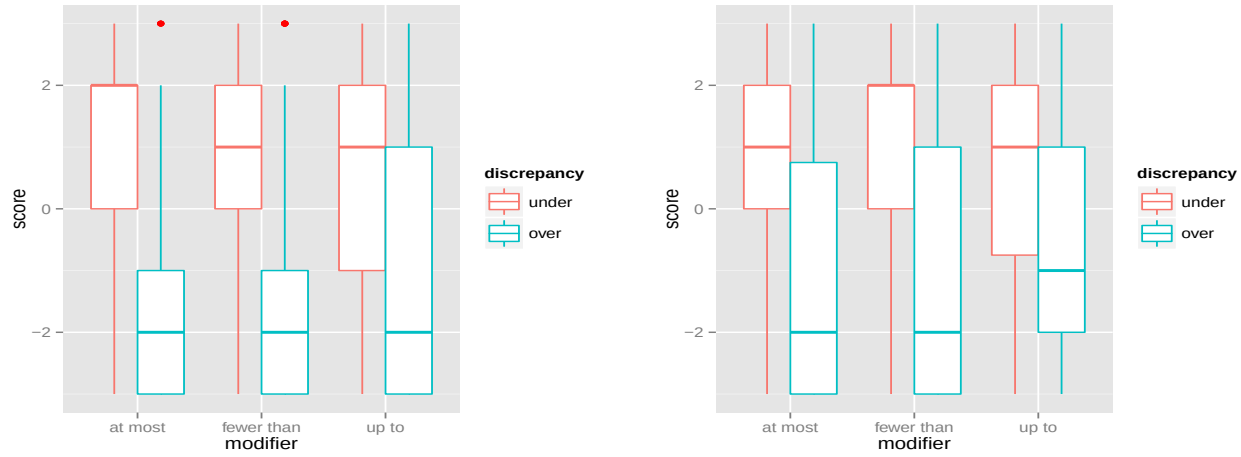


Figure 1: Coherence scores per numeral modifier in the 'under' and 'over' conditions

3.5. Discussion

We assume that what guided participants in rating the coherence of the sentence continuations they read is whether the information in the second and more specific underlined sentence was compatible with the information in the first and more general sentence. Since we manipulated the numbers in the second sentence only, we expect participants to identify the numbers as the crucial point of comparison between the two sentences. The explicit use of anaphora in the second, underlined sentence in eight out of twelve items allows us to assume that participants interpreted the second sentence as stating a specific case included in the general statement in the first sentence. In the remainder third of the items, however, we can't exclude the possibility that participants

interpreted the underlined continuation as an exception rather than a specific case consistent with the more general claim. In this case, we would predict a bigger variety of scores for the 'over' condition of all three numeral modifiers. Indeed, the boxplots in Figure 2b below show that the four items that had no anaphoric terms in the second, underlined sentence received a great range of scores in all three numeral modifier conditions in the 'over' condition, which was remarkably greater especially for *to poli* 'at most' and *lighotero apo* 'less than' compared to the items with anaphora, depicted in Figure 2a. Note also that the only difference between the overall scores presented in Figure 1 and those for the items with anaphora in Figure 2a is the drop in scores for the *lighotero apo* 'less than' items in the 'over' condition. If anything, this implies an even stronger difference between *lighotero apo* 'less than' and *mehri* 'up to' when the number in the continuation is greater than the number in the first sentence.



(a) Items with anaphora

(b) Items without anaphora

Figure 2: Coherence scores per numeral modifier in the 'under' and 'over' conditions for items with/without anaphora in the continuation sentence

Hence, the scores for the 'over' conditions show a clear difference between *mehri* 'up to' on the one hand and *to poli* 'at most' and *lighotero apo* 'less than' on the other hand. The consistently low scores for a sentence continuation with *to poli* 'at most' and *lighotero apo* 'less than' strongly suggest that these numeral modifiers specify an uncancellable upper bound and therefore a continuation with a greater number than the modified numeral is considered a "very bad continuation." In contrast, the significantly higher scores for a sentence continuation with *mehri* 'up to' strongly suggest that the upper bound the numeral modifier *mehri* 'up to' specifies is cancellable, a hallmark property of generalized conversational implicature (Grice, 1975). These results are in favour of Blok (2015), who argues that the upper bound specified by *up to* cross-linguistically is derived pragmatically via scalar implicature, whereas the upper bound specified by *at most* and *less/fewer than*, and their counterparts is part of the semantic content.

The scores for the 'under' condition are puzzling: scores for items with *mehri* 'up to' received significantly lower coherence scores than items with *to poli* 'at most' and their difference from *lighotero apo* 'less than' items was borderline significant. If the semantics of *up to* in Blok's cross-linguistic account is correct, then the fact interns in advertisement companies get up to 980 dollars per month is compatible with interns in some of these companies making (only) 950 dollars. But if this semantics is correct, then why is such a sequence of sentences with *mehri* 'up to', see translation in (12), less coherent than the same sequence with *to poli* 'at most' or *lighotero apo* 'less than', translated in (13)?

(12) Interns in advertisement companies get up to 980 dollars per month;
the interns in some of them are paid 950 dollars per month.

(13) Interns in advertisement companies get $\left\{\begin{array}{c}\text{at most}\\ \text{less than}\end{array}\right\}$ 980 dollars per month;
the interns in some of them are paid 950 dollars per month.

Our hypothesis is that this difference between these numeral modifiers has to do with their monotonicity and how it relates to the expected continuation in discourse. Recall from the discussion in Section 2.4 that both *at most* and *less/fewer than* are monotone decreasing. In contrast with these two numeral modifiers, Blok (2015) argues that *up to* and its counterparts assert a lower-bound, which leads to a monotone-*increasing* semantics. In addition to the entailment patterns of the modifiers, there seems to be some distinct inference a listener would make regarding the expectation of the interlocutors given the use of one of these numeral modifiers that correlates with monotonicity: Upward monotonicity is correlated with the expectation that higher amounts be paid to interns and downward monotonicity is correlated with the expectation that lower amounts be paid to interns (Nouwen, 2006). The use of *mehri* 'up to', by virtue of it being upward monotone, leads participants to construct an expectation in which higher payments for interns is what is expected. Given this expectation, a lower amount as an example of one of the cases is incoherent, as it goes in the opposite direction of that expectation—even though lower amounts (above 0) are entailed. The opposite can be said of *to poli* 'at most' or *lighotero apo* 'less than', in which a constructed question under discussion would be that lower payments are expected. Given that expectation, a lower amount as an example of a specific case is coherent.

There is an alternative interpretation for the lower coherence of follow-up statements with lower numbers after statements with *mehri* 'up to', pointed out to us by Brian Buccola (p.c.). Recall that *up to n* asserts the existence of values between some implicit, contextually-salient lower bound above 0 and *n*. Since the number modified by *up to* is above a certain contextually-salient standard, the resulting inference is that this number is considerably or notably high (for the entities counted). Given the focus on the notable height of the number in question, it will be difficult to find a coherent connection to a subsequent sentence that mentions a specific case in which the number was lower.

Note, however, that similarly to the ability to comment on the evaluative adjective *nice* in (14), a speaker can contest the evaluative component of *up to three* in (15), but can't do the same when (15a) includes *at most three* or possibly even just *three* instead. (Cf. *three whole displays*, which does license (15b).)

(14) a. The CEO is very nice.
b. No she's not! She's condescending and impatient.

(15) a. With the Mac Pro, you can connect $\left\{ \begin{array}{c} \text{\# at most} \\ \text{up to} \\ \text{? –} \end{array} \right\}$ three 4K displays at once.
b. Pfff, three is not such a large number of displays. With a DisplayLink adapter you can connect six monitors to you PC or Mac.

We leave the characteristics of evaluativity evoked by some modifier numerals to future research. In what follows, however, we address the confound introduced by the sentence continuation structure in the Greek stimuli and examine additional factors that may affect the likelihood of upper-bound construals in a follow-up experiment.

4. Experiment 2: English

4.1. Motivation

The findings of Experiment 1 are compatible with Blok's (2015) claim that the upper-bound construal of *up to* cross-linguistically is pragmatically-derived, while *at most* and *less/fewer than*'s is part of the semantic content. In Experiment 2, we make a few modifications. First, we avoid the confound introduced by sentence continuations that lead to a mismatch in expectations by presenting two independent statements which participants have to rate the compatibility of. Second, we examined the effect of scalar distance on the likelihood of upper-bound inferences. Previous studies of scalar implicatures show that a greater distance between alternate values on the scale leads to stronger implicature (Beltrama and Xiang, 2013; van Tiel et al., 2016). If *up to* leads to scalar implicature, then we would expect a similar effect on its strength. Third, we systematically controlled for the roundness of the modified numerals.

4.2. Methods

In the experimental task, we asked subjects to rate to what extent a CLAIM was compatible with a subsequently-provided FACT on a modified Likert scale of -3 to 3, where -3 was "completely

incompatible" and 3 was "completely compatible," as illustrated in the example stimulus in (16). The first statement, the CLAIM, was adapted from naturally occurring tokens gleaned from COCA (Davies, 2008). The CLAIM included one of the three numeral modifiers under investigation and a non-round number. The FACT was a more specific statement about one of the cases the CLAIM was referring to, and included a number as well, which was slightly smaller than the number in the CLAIM ($m_{fact} = n_{claim} * 0.95$; the 'under' condition), much smaller than the number in the CLAIM ($m_{fact} = n_{claim} * 0.25$; the 'way under' condition), slightly greater than the number in the CLAIM ($m_{fact} = n_{claim} * 1.05$; the 'over' condition), or much greater than the number in the CLAIM ($m_{fact} = n_{claim} * 1.75$; the 'way over' condition).

(16) Example target item:

CLAIM: Clarendon High School used its smart classrooms 50 times last year with $\left\{\begin{array}{c}\text{fewer than}\\ \text{at most}\\ \text{up to}\end{array}\right\}$ 39 students participating in this classroom environment.

FACT: On one occasion, the smart classroom was used at Clarendon High School last year, $\left\{\begin{array}{c}10\\ 37\\ 41\\ 68\end{array}\right\}$ students participated.

How compatible is the CLAIM with the FACT?

–3	–2	–1	0	1	2	3
completely incompatible						completely compatible

All target items (N=30) were rotated through 15 lists in a latin square design, so that each participant only saw one condition per item. 30 filler items including 10 contradictions, 10 entailments and 10 scalar implicatures involving a quantifier were added into the mix, too. In the contradiction items, a quantifier like *a couple of* in the CLAIM, as in (17), was contrasted with a quantifier *none* in the FACT, as in (17a). In the entailment items it was contrasted with the quantifier *some* in the FACT, see (17b). And in the implicature items, it was contrasted with the quantifier *all*, as in (17c).

(17) Example of filler items:
CLAIM: The New York Daily News reports that **a couple of** the of the pill bottles were not labeled and were probably older prescriptions.

a. Contradiction:
FACT: **None** of the pill bottles were older prescriptions.
b. Entailment:
FACT: **Some** of the pill bottles were older prescriptions.

c. Implicature:
FACT: **All** of the pill bottles were older prescriptions.

90 declared native speakers of English participated in the experiment on Amazon's Mechanical Turk. Data from six participants were discarded, because they gave scores greater than -1 to the contradiction items and smaller than 1 to the entailment items. Data from the remaining 84 participants (58 female participants; mean age: 38.73; age range: 21–54) were used to the statistical analyses reported on here.

4.3. Results

Similarly to Experiment 1, the data were analyzed with mixed-effects ordered probit regression models using the `ordinal R` package (Christensen, 2013). The full model with type of modified numeral, discrepancy, and their interaction as fixed effects with random intercepts and slopes for modifier, discrepancy, and their interaction for subjects, and with random intercepts and slope for items was found to be the best fit for our data ($LRstatistic = 43.61$, $df = 14$, $p < .001$). Setting LESS/FEWER THAN and WAY UNDER as reference levels, we found that overall scores for items with *up to* were significantly different from items with *less/fewer than* ($\beta = 2.43$, $SE = .20$, $p < .01$) but we found no significant difference between items with *at most* and items with *less/fewer than* ($\beta = .26$, $SE = .20$, $p = .16$). In addition, we found a significant difference between the 'way under' condition and the 'over' ($\beta = -6.61$, $SE = .25$, $p < .0001$) and 'way over' ($\beta = -7.09$, $SE = .25$, $p < .0001$) condition, and marginally significant difference between the 'way under' and 'under' conditions ($\beta = .34$, $SE = .19$, $p = .08$). We found a significant interaction for *up to* and 'over' ($\beta = 2.41$, $SE = .28$, $p < .001$) and marginally significant interaction for *up to* and 'way over' ($\beta = .50$, $SE = .27$, $p = .006$).

We further explored these effects with four additional models, each consisting of one of the four discrepancy conditions ('way under', 'under', 'over,' and 'way over'). All four models had modifier as the only fixed effect. Similarly to Experiment 1, in the 'over' condition, items with *up to* received significantly higher compatibility rates than items with *less/fewer than* ($\beta = 2.03$, $SE = .20$, $p < .01$) and items with *at most* ($\beta = 2.25$, $SE = .20$, $p < .01$). In the 'way over' condition, items with *up to* received significantly higher compatibility rates than items with *less/fewer than* ($\beta = .49$, $SE = .186$, $p < .01$ and items with *at most* ($\beta = .51$, $SE = .18$, $p = .01$).

Finally, we ran three additional models, each consisting of one of the three numeral modifiers (*less/fewer than, at most* and *up to*). For each modifier, items in the 'way over' condition were significantly less coherent than items in the 'over' condition (*less/fewer than*: $\beta = -.46$, $SE = .18$, $p < .05$; *at most*: $\beta = -.95$, $SE = .37$, $p < .01$; *up to*: $\beta = -2.18$, $SE = .26$, $p < .01$). There were no differences between the ratings for the 'under' and 'way under' conditions for any of the modifiers and among them.

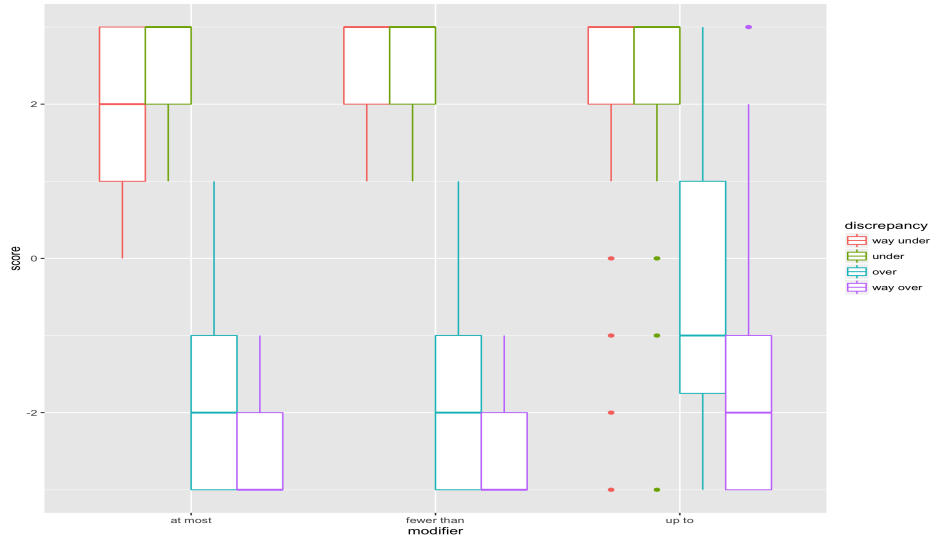


Figure 3: Compatibility scores per numeral modifier in Experiment 2

4.4. Discussion

In Experiment 2, too, a clear difference between *up to* on the one hand and *at most* and *less/fewer than* on the other hand in the 'over' condition strongly suggests that *up to*'s upper bound is far more cancellable, while *at most* and *less/fewer than*'s upper bound is hard to cancel. This is in line with Blok (2015), according to which the upper bound of *up to* is derived via a pragmatic mechanism, while the upper bound of *at most* is derived as an entailment. The small range of scores *at most* received as compared to the large range of scores for *up to* further strengthens this contrast.

Given that the upper bound imposed by *at most* and *less/fewer than* seems to be semantic, we would expect it to be impervious to contextual factors such as the scalar distance of the value above the number specified by *at most/less than n*, in contrast with the pragmatically-derived upper bound communicated by *up to*, which would be more sensitive to contextual factors, similarly to other types of scalar implicature (Doran et al., 2009; Degen, 2015). While we indeed found that greater values in the FACT led to significantly lower rates in CLAIMs with *up to*, we were surprised to find the same difference for CLAIMs with *at most* and for CLAIMs with *less/fewer than*.

The source of the effect between the 'over' and 'way over' conditions can either be the same for all numeral modifiers or be different for *at most* and *less/fewer than* on the one hand and *up to* on the other. If the source of the effect is the same for all numeral modifiers, then what is at play here isn't contextual factors, as the upper bound imposed by *at most* and *less/fewer than* is semantic. What is possibly the case here is that participants mapped the numeral scale onto the Likert scale used in the task. That is, the numbers in the CLAIM both in the 'over' and 'way over' condition were

equally incompatible with the number in the FACT, but the greater distance between the numbers in the 'way over' condition led participants to mark the numerical difference irrespective of the fact that the FACTs in both the 'over' and 'way over' condition were false.

Another possibility is that a great difference between the value in the CLAIM vs. the one in the FACT seems to violate Relevance: If I know of a case in which $\leq$ 68 students participated in the smart classroom (see FACT in (18)), why would I utter a general claim with a much smaller number (e.g., 39 in CLAIM below), let alone a number that points to a fine level of granularity with no pragmatic slack to allow for a larger range of possible values (cf. 40)?

(18) CLAIM: Clarendon High School used its smart classrooms 50 times last year with $\left\{\begin{array}{c}\text{fewer than}\\ \text{at most}\\ \text{up to}\end{array}\right\}$ 39 students participating in this classroom environment.
FACT: On one occasion, the smart classroom was used at Clarendon High School last year, 68 students participated.

In sum, the lower rates for the 'way over' condition could be driven by the violation of Relevance. Although the compatibility rates decrease for all numeral modifiers, the combination of different semantic and pragmatic factors lead to what seems like a similar behaviour. Future studies would be required to tease apart the role of the numeral scale, the scalar distance, granularity, and relevance on the inferences speakers draw from utterances with numeral modifiers.

5. Conclusion

Taken together, the results from the two experiments show that upper-bound construals are more likely in superlative modifiers, like *at most* and *to poli* 'at most', and comparative modifiers, like *less/fewer than* and *lighotero apo* 'less than', than they are for directional modifiers, like *up to* and *mehri* 'up to', suggesting that this contrast is due to the difference in how the upper bound is derived: in superlative and comparative numeral modifiers it is derived from the lexical semantics, whereas in *up to* it is derived as a pragmatic inference, supporting Blok's (2015) account.

In the second experiment we show that the upper-bound implicature is sensitive to an additional contextual factor, namely the scalar distance between possible alternatives and the number modified and asserted. This ties in with previous theoretical and experimental studies that show that the distance of alternatives on an entailment-based scale affects the likelihood of an upper-bound construal (Horn, 1972; Beltrama and Xiang, 2013; van Tiel et al., 2016).

References

Beltrama, A. and M. Xiang (2013). Is good better than excellent? An experimental investigation on scalar implicatures and gradable adjectives. In E. Chemla, V. Homer, and G. Winterstein (Eds.), *Sinn und Bedeutung 17*, pp. 81–98.

Blok, D. (2015). The semantics and pragmatics of directional numeral modifiers. In S. D'Antonio, M. Moroney, and C. R. Little (Eds.), *Proceedings of Semantics and Linguistic Theory 25*.

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

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

Büring, D. (2008). The least *at least* can do. In C. B. Chang and H. J. Haynie (Eds.), *Proceedings of the 26th West Coast Conference on Formal Linguistics*, Somerville, MA, pp. 114–120. Cascadilla Proceedings Project.

Christensen, R. H. B. (2013). Ordinal—regression models for ordinal data, R package version 2013.9-30.

Cohen, A. and M. Krifka (2014). Superlative quantifiers and meta-speech acts. *Linguistics and Philosophy 37*(1), 41–90.

Cummins, C. and N. Katsos (2010). Comparative and superlative quantifiers: Pragmatic effects of comparison type. *Journal of Semantics 27*(3), 271 –305.

Cummins, C., U. Sauerland, and S. Solt (2012). Granularity and scalar implicature in numerical expressions. *Linguistics and Philosophy 35*(2), 135–169.

Davies, M. (2008). The corpus of contemporary American English: 425 million words, 1990-present.

Degen, J. (2015). Investigating the distribution of some (but not all) implicatures using corpora and web-based methods. *Semantics and Pragmatics 8*(11), 1–55.

Doran, R., R. M. Baker, Y. McNabb, M. Larson, and G. Ward (2009). On the non-unified nature of scalar implicature: An empirical investigation. *International Review of Pragmatics 1*(2), 211–248.

Fox, D. and M. Hackl (2006). The universal density of measurement. *Linguistics and Philosophy 29*(5), 537–586.

Gazdar, G. (1979). *Pragmatics: Implicature, presupposition, and logical form*. Academic Press New York.

Geurts, B. (2006). Take 'five': the meaning and use of a number word. In L. Tasmowski and S. Vogeleer (Eds.), *Indefiniteness and plurality*. Amsterdam: Benjamins.

Geurts, B. (2010). *Quantity Implicatures*. Cambridge University Press.

Geurts, B. and R. Nouwen (2007). 'At least' et al.: The semantics of scalar modifiers. *Language 83*(3), 533–559.

Grice, H. P. (1975). Logic and conversation. In P. Cole and J. Morgan (Eds.), *Syntax and Semantics*, Volume 3: Speech Acts, pp. 41–58. New York: Academic Press.

Hackl, M. (2000). *Comparative Quantifiers*. Ph. D. thesis, MIT.

Horn, L. (1992). The said and the unsaid. In C. Barker and D. Dowty (Eds.), *SALT II: Proceedings of the Second Conference on Semantics and Linguistic Theory*, Number 40 in The Ohio State University Working Papers in Linguistics, Columbus, pp. 163–192. Ohio State University.

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

Huang, Y. T., E. Spelke, and J. Snedeker (2013). What exactly do numbers mean? *Language Learning and Development 9*(2), 105–129.

Kennedy, C. (2013). A scalar semantics for scalar readings of number words. In I. Caponigro and C. Cecchetto (Eds.), *From grammar to meaning: The spontaneous logicality of language*. Cambridge University Press.

Kennedy, C. (2015). A "de-Fregean" semantics (and neo-Gricean pragmatics) for modified and unmodified numerals. *Semantics and Pragmatics 8*, 10–1.

Krifka, M. (1999). At least some determiners aren't determiners. In K. Turner (Ed.), *The Semantics/Pragmatics Interface from different points of view*, pp. 257–291. Oxford: Elseviewer Science.

Krifka, M. (2007). Approximate interpretation of number words: A case for strategic communication. In G. Bouma, I. Kräer, and J. Zwarts (Eds.), *Cognitive Foundations of Communication*, pp. 111–126. Amsterdam: Koninklijke Nederlandse Akademie van Wetenschapen.

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

Mayr, C. (2013). Implicatures of modified numerals. In I. Caponigro and C. Cecchetto (Eds.), *From grammar to meaning: The spontaneous logicality of language*, pp. 139–71. Cambridge University Press.

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

Nouwen, R. (2006). Remarks on the polar orientation of *almost*. In J. van de Weijer and B. Los (Eds.), *Linguistics in the Netherlands*, Volume 23, pp. 162–173. AvT Publications, John Benjamins Publishing Company.

Nouwen, R. (2010). Two kinds of modified numerals. *Semantics and Pragmatics 3*, 1–41.

Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy 27*(3), 367–391.

Schwarz, B. (2011). Remarks on class B numeral modifiers. Handout of a talk at the workshop *Indefinites and Beyond*, Universität Göttingen, November 19, 2011.

Schwarz, B. (2013). 'At least' and quantity implicature: Choices and consequences. In M. Aloni, M. Franke, and F. Roelofsen (Eds.), *Proceedings of the 19th Amsterdam Colloquium*, pp. 187–194.

Schwarz, B., B. Buccola, and M. Hamilton (2012). Two types of class B numeral modifiers: A reply to Nouwen 2010. *Semantics and Pragmatics 5*, 1–25.

van Tiel, B., E. Van Miltenburg, N. Zevakhina, and B. Geurts (2016). Scalar diversity. *Journal of Semantics 33*(1), 137–175.