Proceedings of SALT 24: 000–000, 2014

Ignorance in context: The interaction of modified numerals and QUDs *

Matthijs Westera
Institute for Logic, Language and Computation, University of Amsterdam

Adrian Brasoveanu
University of California Santa Cruz

Abstract We argue for a purely pragmatic account of the ignorance inferences associated with superlative but not comparative modifiers (*at least* vs. *more than*). Ignorance inferences for both modifiers are triggered when the question under discussion (QUD) requires an exact answer, but when these modifiers are used out of the blue the QUD is implicitly reconstructed based on the way these modifiers are typically used, and on the fact that *at least n*, but not *more than n*, mentions and does not exclude the lower bound *exactly n*. The paper presents new experimental evidence for the context-sensitivity of ignorance inferences, and also for the hypothesis that the higher processing cost reported in the literature for superlative modifiers is context-dependent in the exact same way.

Keywords: superlative vs. comparative modifiers, ignorance inferences, questions under discussion, experimental semantics and pragmatics

1 Introduction

Consider the brief exchange in (1) below. Given that B uses the superlative modifier *at least*, we infer that B was unable to give a more precise answer:

(1) A: Exactly how many students took Experimental Pragmatics?
B: At least ten students took Experimental Pragmatics.
Inference: B doesn't know the exact number of students.

Let us call this an ignorance inference. Ideally, one would explain this inference pragmatically, along the lines of Grice (1989): if B had known a more complete answer to A's question, she (being cooperative) would have said so; but B used the modifier *at least*, whose role is to indicate that the number corresponds to a lower bound and not to the exact quantity that A inquired about. We therefore infer that B doesn't know that exact quantity.

* We want to thank several anonymous SALT 24 reviewers and the SALT 24 participants for comments and discussion. The usual disclaimers apply.

Although this would be the most parsimonious account, it is challenged by the strong intuition reported in Nouwen (2010) that out of the blue, the superlative modifier *at least* triggers an ignorance inference while the comparative modifier *more than* does not:[1]

(2) *[When uttered by someone with basic knowledge of geometry]*

a. ? A hexagon has at least five sides.

b. A hexagon has more than four sides.

The pragmatic reasoning sketched above should work for both types of modifiers alike, suggesting that pragmatics alone is insufficient to capture the pattern of ignorance inferences exhibited by these two kinds of scalar modifiers. Consequently, existing approaches attribute the ignorance inference to a lexical property specific to *at least* rather than to general pragmatics: *at least* is in some sense disjunctive (Büring 2008; Cummins & Katsos 2010; Coppock & Brochhagen 2013b).

In this paper, we argue that these approaches deviate unnecessarily from the more parsimonious pragmatic account. Under this account, ignorance inferences for both *at least* and *more than* depend primarily on whether a precise answer was required or not, i.e., on the question under discussion (QUD). The contrast in (2), we argue, arises only because the context is underspecified, and we are left to guess what the QUD was like. When guessing, any property of *at least* or *more than* can affect our judgments, in particular: (*i*) the way *at least n* and *more than n* are typically used, and (*ii*) the fact that *at least n*, but not *more than n*, mentions and does not exclude the lower bound *exactly n* (Nouwen 2010; Cummins & Katsos 2010; Coppock & Brochhagen 2013a, among others). We argue that (either of) these two differences are sufficient to account for the contrast in (2), so there is no reason to stipulate deeper lexical differences between superlative and comparative modifiers only to explain the difference in ignorance inferences between them.

The second contribution of this paper is empirical: we present new experimental evidence for the context-sensitivity of ignorance inferences, and therefore for analyzing the contrast in (2) in terms of contextual underspecification.

Finally, reading time data from one of our experiments suggests that the higher processing cost reported in the literature for superlative modifiers (e.g.,Geurts et al. 2010; Cummins & Katsos 2010) is in fact context-dependent in the exact same way.

The paper is structured as follows. Section §2 presents the details of the proposed explanation, including a small corpus study corroborating one of its assumptions. Section §3 presents new experimental evidence that supports the predicted context dependence of ignorance inferences, as well as our claim that contextual underspecification is the reason for the contrast in (2). Section §4 briefly reviews previous

1 This contrast has been corroborated experimentally, see Geurts, Katsos, Cummins, Moons & Noordman (2010); Cummins & Katsos (2010).

experimental results and compares our account to existing proposals, and section §5 concludes.

2 The context dependence of ignorance inferences

As already mentioned, the ignorance inference in (1) can be explained pragmatically: A asked for a precise answer, B didn't give one,[2] hence B wasn't able to (assuming B is cooperative).[3] This kind of pragmatic explanation makes two predictions that are of particular interest. First, it predicts that ignorance inferences arise only if a precise answer was asked for. In particular, we do not expect ignorance inferences if the same utterance as in (1) above answers a polar question:

(3) **Predicted**: no ignorance inference.
A: Did at least ten students take Experimental Pragmatics?
B: At least ten students took Experimental Pragmatics.

The second prediction of interest is that ignorance inferences do not depend on the type of modifier used in the answer (superlative vs. comparative): B's response fails to give an exact answer regardless. Thus, we expect the same pattern for *more than*:

(4) **Predicted**: ignorance inference.
A: Exactly how many students took Experimental Pragmatics?
B: More than ten students took Experimental Pragmatics.

(5) **Predicted**: no ignorance inference.
A: Did more than ten students take Experimental Pragmatics?
B: More than ten students took Experimental Pragmatics.

An experimental investigation of these predictions is presented in section §3 below. The remainder of this section discusses how this line of explanation accounts for the contrast between superlative and comparative numerals when they are used out of the blue (2), and briefly compares this account to existing work.

2.1 Dealing with contextual underspecification

What happens if the relevant parts of the context needed to pragmatically compute ignorance inferences are left implicit? Consider the example below, similar to (2):

2 We assume that the presence of numeral modifier in B's utterance indicates that the answer is merely partial and should not receive an exhaustive/exact interpretation. It is not necessary for our purposes to spell out the nature of this partial answerhood in more detail. See Cummins, Sauerland & Solt (2012) for more discussion, and in particular for the fact that numeral modifiers can trigger inferences to the effect that higher numerals of the same granularity are to be excluded.

3 This is the standard story of how 'Quantity implicatures' are explained, but we will call them inferences rather than implicatures because we don't want to commit to the idea that the ignorance inference is always part of what the speaker meant.

(6) a. At least ten students took Experimental Pragmatics. *Ignorance:* ✓

b. More than ten students took Experimental Pragmatics. *Ignorance:* *

Following Kadmon & Roberts (1986) among others, we assume that when the context is underspecified, the audience must guess what the context is like – and simpler/typical contexts are more likely to be imagined than more complex/atypical contexts.[4] We therefore propose that the contrast in (6) is due to a difference in the typical context of use for *at least* vs. *more than*, with the former being used more typically in contexts demanding a precise answer.[5]

Note that there are two other ways to explain this contrast. First, a completely non-pragmatic audience might ignore the context dependence and instead look directly at whether the sentences typically convey ignorance or not. Second, a maximally pragmatic audience might not care about typical usage and would look instead at the properties of *at least* vs. *more than* that caused those differences in usage to begin with. These three strategies might be behaviorally indistinguishable from each other. Our choice to pursue an explanation in terms of guessing the context, which falls in between the other two, is primarily a rhetorical one: it most clearly singles out contextual underspecification as the reason for the contrast in (2) and (6), which connects naturally to the experiments reported in section §3.

2.2 Quantifying the context of use for *at least* vs. *more than*

We did a small corpus study to test if the typical contexts of use for *at least* are 'precise contexts' more often than for *more than*. Crucially, we took round numbers to be indicators of imprecise/approximate contexts, following Jansen & Pollmann (2001), Krifka (2009) and Cummins (2011) among others.

If *at least* is more typically used than *more than* in contexts where an exact answer is called for, one would expect the distribution of *at least* to be less sensitive than the distribution of *more than* to whether the modified numeral is round or

4 Kadmon & Roberts (1986) highlight especially the importance of intonation as a cue for what the QUD is like (see also Roberts 1996). Indeed, it seems to us that B's responses in (1) and (4) (with ignorance) may come with an accent on the modifier and a high final boundary tone to signal the incompleteness, while (3) and (5) (without ignorance) seem to be typical 'broad focus' sentences with a falling pitch. Since our experiments and corpus study did not examine intonation, we discuss the role of contextual underspecification abstracting away from any intonational cues. It is expected, however, that the contrast in (2) and (6) occurs only if both context and intonation are underspecified.

5 Mere differences in typicality can give rise to stark categorical contrasts if the participants in the experiment are forced to choose, especially if they assume the experimenter is cooperative (as argued by Schwarz 1996): the participants conjecture that if the context had been an atypical one, a cooperative experimenter would have made it explicit. However, our experiments were not designed to inform us about this possibility, so we will set it aside.

not. And this is what we found in the Corpus of Contemporary American English (COCA; 450 million words, Davies 2008-).

For each numeral $n \in \{1, \ldots, 100\}$, we obtained the counts of n (mean: 28492), *at least n* (mean: 123) and *more than n* (mean: 317).[6] We used Poisson regression (log-linear) models with overdispersion to analyze this count data.[7] In our models, we controlled for the differences in frequency induced by the magnitude of the numeral (i.e., 1 through 100) by including all the polynomial terms for this predictor of degree ≤ 4. We were interested in the interaction of numeral form – BARE NUMERAL VS. AT LEAST VS. MORE THAN, with BARE NUMERAL as the reference level – and roundness. For roundness, we separately considered divisibility by 5 (NOT ROUND VS. ROUND5) and by 10 (NOT ROUND VS. ROUND10), with NOT ROUND the reference level in both cases.

The interaction of MORE THAN and ROUND5 was positive and highly significant ($\beta = 2.17, SE = 0.49, p = 1.2 \times 10^{-5}$), while the interaction of AT LEAST and ROUND5 was only borderline significant ($\beta = 1.14, SE = 0.64, p = 0.08$). We obtained similar results for interaction of MORE THAN and ROUND10 ($\beta = 2.22, SE = 0.34, p = 3.9 \times 10^{-10}$) and the interaction of AT LEAST and ROUND10 ($\beta = 1.14, SE = 0.52, p = 0.03$). These results show that both modified numerals are used more frequently with round numbers than the corresponding bare numberals, but this increase in frequency is much higher for *more than* than for *at least*. The difference between the two modifiers is highly significant both for ROUND5 and ROUND10, as determined by a direct comparison of MORE THAN and AT LEAST).

The upshot is that the frequency of *at least n* is significantly less sensitive to whether n is round than the frequency of *more than n*. We take this to indicate that *at least* is used more frequently than *more than* in precise contexts. Hence, if the audience is left to guess the context in which *at least* is uttered, they will be more likely to guess that the context is precise (compared to what they would guess for *more than*), as required for our explanation of the contrast in (2) and (6).

2.3 Explaining the difference between *at least* and *more than*

A plausible reason for the observed pattern is the fact that *at least n*, but not *more than n* intuitively mentions but does not exclude the *exactly n* possibility. This difference was observed by Cummins & Katsos (2010), who propose that *at least* draws

6 The search strings were 1|2|3|4|...|100, *at least* 1|2|3|4|...|100 and *more than* 1|2|3|4|...|100, where | is disjunction in the query language. The results were obtained on July 18, 2014. We also did a search in which the Arabic numerals were replaced by the corresponding English numerals, and the results are very similar but slightly weaker because of the overall lower counts.

7 All statistical modeling reported in this paper was performed in R (R Core Team 2014).

attention to the numeral and the 'possibility of equality'.[8] Coppock & Brochhagen (2013a) say that *at least* 'highlights' a possibility, a notion borrowed from Roelofsen & van Gool (2010). Similarly, Nouwen (2010) credits a reviewer for the suggestion to identify pragmatic consequences of the fact that *at most n* includes the possibility of *n* while *less than n* excludes it.[9] This difference notwithstanding, Nouwen remarks that it is unclear how this kind of difference could be used to account for the contrast in (2)/(6). Similarly, Coppock & Brochhagen do not use highlighting to that end.

For our purposes, however, this difference between *at least* and *more than* is enough. We don't need it to explain a contrast in ignorance inferences, but merely to explain a tendency for *at least* to be used more frequently in precise contexts. Let us assume that a cooperative speaker would not highlight any particular possibility unless that particular possibility is relevant, and that the typical way for something to be relevant is for it to constitute a possible answer to the QUD (see Roberts 1996 for this way of spelling out the notion of relevance). It follows that using *at least* in a context where no precise answer is required is less preferred, as confirmed by our corpus study, and as required for our explanation of the contrast in (2)/(6).[10]

3 An experimental investigation

The main predictions of the account outlined so far are that (*i*) if the context is sufficiently fixed, ignorance inferences are purely context-dependent (and independent of the modifier type), and (*ii*) if the context is insufficiently fixed, a contrast between the two modifier types will arise. To test these predictions, we performed two experiments with the same overall design, but partly different conditions.

Experimental method. In both experiments, participants were presented with conversations taking place in a courtroom between a judge and a witness. Each item consisted of a question asked by the judge, the answer given by the witness, and a conclusion drawn by the judge based on the conversation. The three components – question, answer and conclusion – were presented in that order in three subsequent screens, as illustrated in figure 1. The crucial manipulations were the type of question posed by the judge (QUD) and the type of scalar modifier in the witness's answer (MOD). The conclusion drawn by the judge in all items was

8 They note this generalizes to other Type B modifiers, e.g., *not more than 20*, *20 or more*.

9 Note that one does not need to stipulate a special meaning component for *at least* to achieve this: highlighting is something *at least n*, but not *more than n*, achieves as a consequence of its truth-conditional meaning (not excluding *exactly n*) combined with its surface form (containing the numeral *n*). We leave a more detailed investigation of this contrast for a future occasion.

10 As mentioned, we cannot rule out that an audience is directly sensitive to this difference between the modifiers, rather than to the difference in typical use it causes.

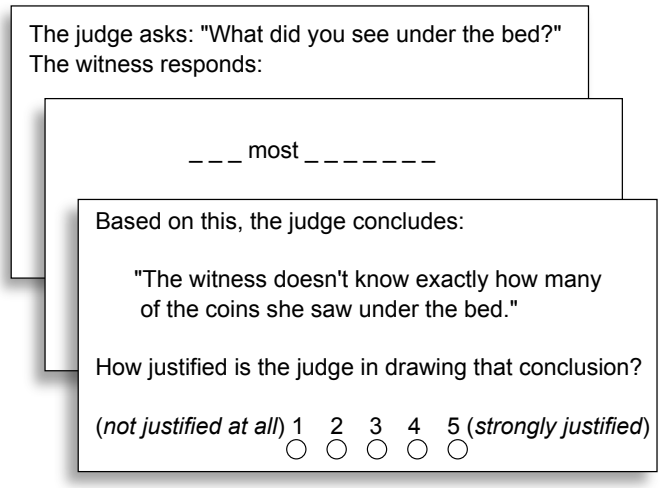

Figure 1 The three components of each item were presented on three separate screens, depicted in order from back to front. The self-paced reading task was on the second screen and the acceptability/validity-judgment task on the third.

an ignorance inference. The experiments combined a self-paced reading and an acceptability/validity-judgment task. First, reading the witness's answer (on the second screen) involved a self-paced reading task: participants read it word-by-word, with the SPACE bar revealing the next word and hiding the preceding one, and we recorded reading times per word. Second, to measure how justified the participants found the judge's conclusion (on the third screen), participants were asked to indicate this on a 5-point Likert scale (1: not justified at all, 5: strongly justified).

In both experiments, the factor QUD had three levels and MOD two, for a total of six conditions. In Experiment 1, QUD ranged over {POLAR, WHAT, HOWMANY} (reference level: POLAR). In Experiment 2, QUD ranged over {APPROX, EXACT, DISJUNCT} (reference level: APPROX). In both experiments, MOD ranged over {COMP, SUP} (reference level: COMP). Example items are provided below. The underlined content words varied in non-crucial ways across items, with verbs in the set {*find*, *see*, *hear*}, and a variety of nouns and prepositional phrases. In POLAR, the numeral modifier always corresponded to that used in the witness's answer.

(7) Judge's question type (QUD) in Experiment 1:

POLAR Did you find {at most / less than} ten of the diamonds under the bed?

WHAT What did you find under the bed?

HOWMANY How many of the diamonds did you find under the bed?

(8) Judge's question type (QUD) in Experiment 2:

APPROX Approximately how many of the diamonds did you find under the bed?

EXACT Exactly how many of the diamonds did you find under the bed?

DISJUNCT Did you find eight, nine, ten, or eleven of the diamonds under the bed?

(9) Witness' answer type (MOD) in both experiments:

SUP I found at most ten of the diamonds under the bed.

COMP I found less than ten of the diamonds under the bed.

(10) Judge's ignorance inference in both experiments:
The witness doesn't know exactly how many of the diamonds she found under the bed.

There were 6 items per condition, for a total of 36 items. We used a Latin square design: in each experiment, we generated 6 lists of items; in each list, the items were rotated through the 6 conditions, with the items balanced across conditions, and every item appearing on the list exactly once. The participants were rotated through the 6 lists. There were 108 stimuli total (36 items + 72 fillers) in each experiment, the order of which was randomized for every participant. Experiment 1 had 35 participants, Experiment 2 had 51 participants, all undergraduate students at UC Santa Cruz that were native speakers of English and that completed the experiment for course (extra) credit.

Situating the dialogues in a courtroom was meant to serve several purposes. First, we wanted to test the exact same answer against different questions, which meant that we could not elide any material from the answers to make it sound more natural. Thus, the witness's answer is always overly explicit. We hoped that a courtroom setting, where many more things need to be made explicit compared to regular conversations, would mitigate this unnaturalness. Second, a courtroom setting allowed us to have prior instructions in which we could explicitly and justifiably say that the witness has nothing to hide and is fully cooperative, so that the judge's conclusions would not be further complicated by the witness possibly lying. Finally, we hoped that a courtroom theme would keep the task interesting for the participants.

Several choices made in the item design require further motivation. First, the modifier type MOD was always downward-entailing, i.e., *at most* or *less than*. The reason is that a previous experiment by Coppock & Brochhagen (2013a) revealed a contrast between downward but not upward entailing modifiers. By testing downward-entailing modifiers, we therefore maximized the chance of finding a contrast, i.e., maximized the chance of falsifying our own hypothesis. Furthermore, the cases in which we didn't find a contrast (to be discussed presently) support Coppock & Brochhagen's thesis that the contrast they found was not due to an ignorance inference. However, when we interpret our results, we will make the common

assumption that as far as ignorance inferences are concerned, superlative modifiers behave alike irrespective of their monotonicity, just as comparative modifiers do.

Second, the verbs ranged over *find*, *see* and *hear*, all perception verbs, because we wanted the witness to be an authority over what she claimed.[11] In this way, we hoped to prevent participants from imagining a scenario in which the judge already knew the true answer, or in which the witness would express (a possibly mistaken) personal opinion/judgment, things we thought might compromise the perceived validity of the judge's conclusions.

Third, the numeral was always *ten* because it was (*i*) small enough to be precisely countable; (*ii*) big enough for uncertainty about the precise amount to be plausible; and (*iii*) round so that it could be naturally used in an imprecise context. We wanted to minimize the inherent bias of the numeral towards precise or coarse contexts, thus maximizing the chance of finding a contrast between *at most* and *less than*.

Finally, the questions, answers and conclusions always contained prepositional phrases. This may have made the items overly explicit, but it was necessary for the self-paced reading task: effects in such tasks are often delayed and occur 2 or more words after the experimental manipulation point. Because we needed prepositional phrases in the witness's answers, they had to be included in the questions and conclusions, lest either the answer be too specific or the conclusion invalid.

The fillers had the same general structure (question, answer, inference). In each experiment, the fillers contained the same types of questions as the items, with one additional question type (*Did you find ten diamonds under the bed?*). The fillers included obviously valid inferences (15 in Experiment 1, 31 in Experiment 2), plausible inferences (20 in Experiment 1, 15 in Experiment 2), implausible inferences (15 in Experiment 1, 5 in Experiment 2), and obviously invalid inferences (17 in Experiment 1, 19 in Experiment 2), as judged by the experimenters. These fillers were used to filter out participants that did not properly complete the experimental task: 1 participant was filtered out from Experiment 1, and 3 from Experiment 2. The final number of participants was 34 in Experiment 1, and 48 in Experiment 2.[12]

11 *Hear* was half as frequent as *see* and *find*.

12 We provide here more details about the fillers. 20 fillers contained the partitive construction *at most ten of the diamonds*, as in the items, and the rest the non-partitive construction *at most ten diamonds*. Most fillers contained an adverb *approximately*, *probably* or *certainly* in the judge's question or the witness's answer. Most fillers contained a numeral modifier, generally upward entailing, *only*, or *nearly*. The fillers in Experiments 1 and 2 differed with respect to the judge's inference. In Experiment 1, the filler inferences were all different from the item inferences, and were of two forms: (*i*) *The witness considers it possible that she saw* $\{9, 10, 11\}$ *of the diamonds under the bed* and (*ii*) *The witness thinks the number of diamonds she saw under the bed is comparably high*. In contrast, the fillers in Experiment 2 were kept the same throughout: they were all ignorance inferences.

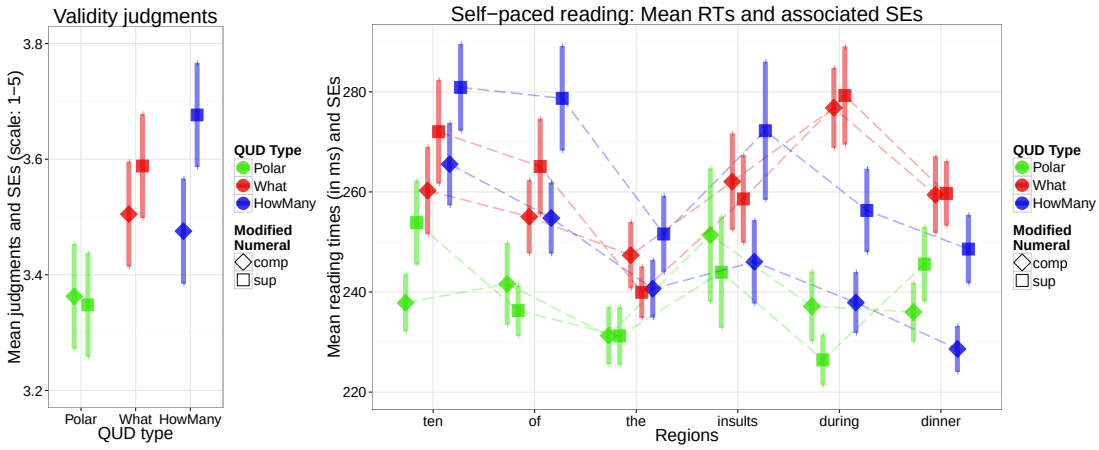

Figure 2 Mean validity judgments, mean RTs and corresponding SEs for Experiment 1.

Results: Validity judgments. Since the validity-judgment data for both experiments was ordinal (ordered categorical), we used mixed-effects ordinal probit regression models to analyze the data. All models included intercept random effects for participants and items.[13]

In Experiment 1, the interaction model (with all two-way interactions) did not significantly reduce deviance compared to the main-effects only model ($LR\ statistic = 2.72, df = 2, p = 0.26$), but when we examined subsets of the data by QUD type, there was a significant effect of SUP for the HOWMANY subset ($\beta = 0.27, SE = 0.11, p = 0.016$), but not for the POLAR and WHAT subsets. In the main-effects only model estimated for the entire data set, the main effects for both WHAT ($\beta = -1.23, SE = 0.08, p = 0.003$) and HOWMANY ($\beta = 0.28, SE = 0.08, p = 0.0004$) were highly significant but not the main effect for SUP, which was only borderline significant ($\beta = 0.11, SE = 0.06, p = 0.08$).

The analysis of the Experiment 2 data proceeded in a very similar way. Once again, the interaction model did not significantly reduce deviance compared to the main-effects only model ($LR\ statistic = 2.2, df = 2, p = 0.33$). Furthermore, SUP was not significant in any of the three QUD subsets.[14] In the main-effects only model estimated for the entire data set, the main effects for both EXACT ($\beta = 0.25, SE = 0.07, p = 0.0001$) and DISJUNCT ($\beta = 0.20, SE = 0.07, p = 0.003$) were highly significant, but not the main effect for SUP ($\beta = 0.08, SE = 0.05, p = 0.14$).

13 The models were estimated using the *ordinal* R package (Christensen 2013).

14 It was borderline significant in the DISJUNCT subset: $\beta = 0.16, SE = 0.09, p = 0.09$.

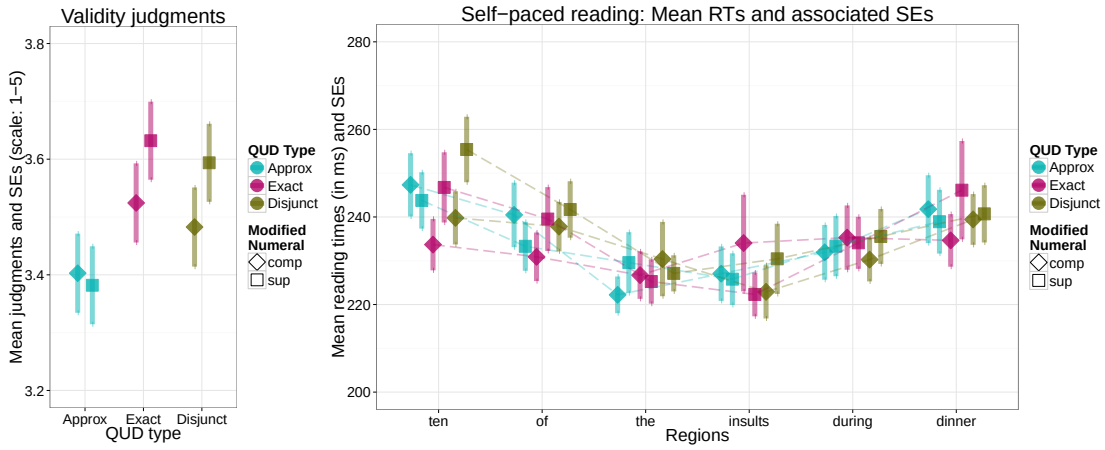

Figure 3 Mean validity judgments, mean RTs and corresponding SEs for Experiment 2.

Results: Reading times. We log-transformed the reading time (RT) data to mitigate its characteristic right skewness. Following Trueswell, Tanenhaus & Garnsey (1994), we factored out the influence of word length and position by running a linear mixed-effects regression with log reading times (log RTs) as the dependent variable. The regression had intercept-only random effects for subjects and two fixed effects: word length (in characters) and word position in the sentence. The resulting residualized log RTs were used for all subsequent analyses.

The analysis focused on the five words following the numeral modifiers. These words, underlined in the example below, were identical across all 6 conditions:[15]

(11) I found at most ten of the diamonds under the bed.

Any effect of modifier type on reading times should be visible in this underlined region. Each word was analyzed using a mixed-effects linear regression model with the residualized log RTs as the dependent variable, the full fixed-effect structure (main effects plus all two-way interactions), and the full random-effects structure for subjects and items that converged (as determined by backwards model selection; see Barr, Levy, Scheepers & Tily 2013 for more discussion).[16]

In Experiment 1, the self-paced reading results are largely congruent with the

15 Word 5 was the ante-penultimate word in the sentence for 35 out of 36 items. In the remaining item, it was the penultimate word. Thus, word 5 was never the wrap-up region (the last word in the sentence).

16 The models were estimated using the *lme4* R package (Bates, Maechler, Bolker & Walker 2014). All the p values were obtained with the *lmerTest* R package (Kuznetsova, Bruun Brockhoff & Haubo Bojesen Christensen 2014). Whenever the *lmerTest* package couldn't estimate p-values, we report the 95% confidence intervals (CIs) estimated by the *lme4* package: we report profile likelihood CIs whenever they could be computed without error; otherwise, we report the (less precise) Wald CIs.

validity judgments: higher RTs are generally correlated with an increase in ignorance inferences. In what follows, we will determine statistical significance by examining the 95% confidence interval (CI) around the coefficient estimates: if the CI excludes 0, the effect is statistically significant. We report the profile likelihood CIs if they can be computed without error, otherwise the (less precise) Wald CIs.

In word 1 (ten in (11) above), the main effect for HOWMANY was positive and significant ($\beta = 0.09, SE = 0.03, CI = (0.03, 0.15)$), and so was the main effect for WHAT ($\beta = 0.07, SE = 0.03, CI = (0.007, 0.13)$). None of the other terms, i.e., neither the main effect for SUP nor any of the two-way interactions, reached significance. Word 2 (of) exhibited the same pattern: only the main effects for HOWMANY and WHAT were positive and significant ($\beta = 0.06, SE = 0.03, CI = (0.004, 0.12)$ and $\beta = 0.06, SE = 0.03, CI = (0.002, 0.11)$, respectively). In word 3 (the), only the main effect for WHAT was significant ($\beta = 0.06, SE = 0.03, CI = (0.009, 0.12)$). No effect was significant in word 4 (diamonds).

But the most interesting region is word 5 (under. There is a significant main effect for WHAT, continuing the trend from the previous words/regions ($\beta = 0.15, SE = 0.04, CI = (0.08, 0.22)$). Importantly, however, we see a positive and significant interaction between HOWMANY and SUP ($\beta = 0.10, SE = 0.045, CI = (0.009, 0.18)$).

When we considered the summed residualized log RTs for the full sentence, the same pattern as in word 5 emerged.[17] The main effect for WHAT was significant ($\beta = 0.6, SE = 0.21, CI = (0.15, 0.99), p = 0.01$) and the interaction between HOWMANY and SUP was borderline significant ($\beta = 0.59, SE = 0.31, CI = (-0.03, 1.20), p = 0.07$), with no other effects reaching significance.

In Experiment 2, none of the effects (main effects or interactions) reached significance on any of the words. The interactions of EXACT and SUP and of DISJUNCT and SUP were numerically positive in the full-sentence RTs, but they did not reach significance.

Summary of main findings and discussion. In Experiment 1, we found an overall stronger ignorance inference in response to a WHAT question than in response to a POLAR question, but with no difference between superlative and comparative modifiers. The contrast in ignorance between SUPs and COMPs is detectable only in responses to HOWMANY questions, with SUPs exhibiting stronger ignorance inferences than COMPs. Stronger ignorance inferences are also systematically correlated with increased RTs: WHAT questions take longer than POLAR questions for both SUPs and COMPs, and the HOWMANY & SUP condition is also associated

17 Three extreme outlier observations out of a total of 1224 were trimmed from the full-sentence data; see Baayen & Milin (2010) for a justification and more discussion of such *post hoc* trimming of reaction-time data.

with higher RTs (but not the HOWMANY & COMP condition).

The higher RTs for WHAT and HOWMANY relative to POLAR QUDs could in principle be due to the fact that in the case of POLAR QUDs, the question already contains the numeral modified in the answer. However, there is no difference between POLAR and HOWMANY QUDs for word 5 (under), as well as for the full-sentence RTs, when a comparative modifier is used. This indicates that the priming effect for POLAR QUDs was not significantly higher than for HOWMANY QUDs. Since priming would have affected comparative and superlative modifiers across the board, the higher RTs we observe with HOWMANY QUDs when a superlative modifier is used must be attributed to the superlative nature of the modifier rather than to a priming effect.

In Experiment 2, we found weaker ignorance inferences for APPROX relative to EXACT and DISJUNCT, with no difference between SUPs and COMPs for any QUD type and no significant effects on RTs. The absence of effects on RTs may be due to habituation: unlike in Experiment 1, all the fillers in Experiment 2 had the judge infer ignorance, just as the experimental items did.

The fact that in both experiments the two types of modifiers behave alike for most QUD types seems to call for an explanation like the one we proposed. According to this explanation, ignorance inferences are primarily context/QUD driven: POLAR and APPROX QUDs ask for a coarse answer, while WHAT, EXACT and DISJUNCT QUDs ask for a precise answer. The experimental results are problematic for an account that connects the ignorance inference of superlative modifiers to anything substantially stronger than a tendency that can be overruled by the context.

Because the explanation predicts that a contrast between the two types of modifiers may occur only when the context is underspecified, the HOWMANY QUD must leave the context partially underspecified. We can think of this as a hidden 'granularity' parameter analogous to covert/implicit quantifier domain restriction.

The classification of QUDs emerging from the experimental results is intuitively plausible, except perhaps for the classification of WHAT QUDs as specific. Note however that in the COCA corpus, *exactly what* outnumbers *approximately what* by about 500 to 1 (likewise if the modifier comes after *what*), far greater than what we would expect based on *exactly how many* and *approximately how many* (about 15 to 1) or *exactly* and *approximately* on their own (about 4 to 1).[18] This suggests that WHAT QUDs are less neutral than HOWMANY, corroborating our experimental results.

Finally, if we look only at Experiment 1, the fact that RTs for *at most* and *less than* differ only if the context is underspecified (as it happens for HOWMANY

18 The following counts were obtained Aug. 15, 2014: *exactly what* – 14217, *approximately what* – 30, *what exactly* – 2154, *what approximately* – 2, *exactly how many* – 314, *approximately how many* – 22, *exactly* – 61632, *approximately* – 17082.

QUDs, for example) suggests that this difference in RTs, just like the difference in validity judgments, should be understood primarily in pragmatic terms. The increased RTs may reflect the cost of the pragmatic ignorance inference, or they may be due to context-dependent intonational effects that may occur during silent reading via subvocalization (Fodor 2002).

4 Relation to existing work

Previous accounts. For Geurts & Nouwen (2007), *at least 5* means certainly 5 and possibly more. This is often cited as a 'semantic' account of the ignorance inference, but strictly speaking they don't say anything about the ignorance inference. For instance, *at least 5* is semantically true even if the speaker knows there are exactly 6. Geurts et al. (2010: 134) realize this and propose somewhat implicitly that ignorance is implied pragmatically, in the same way that *possible* may imply *not certain* (presumably via the Maxim of Quantity). But they don't spell this out in any detail.

Geurts & Nouwen's approach was criticized by Cummins & Katsos (2010), for instance for predicting the wrong truth conditions for *at most* in the antecedent of a conditional (also noted by Geurts & Nouwen 2007). Cummins & Katsos favor instead a pragmatic account along the lines of Büring (2008). Büring takes *at least 5* to be like the disjunction *5 or more*, and argues that disjunction-like meanings typically convey ignorance in virtue of (a special instance of) the Maxim of Quantity that says: since the speaker used a disjunction (or an expression with a disjunction-like meaning), she must be ignorant about the individual disjuncts. Büring already notes that it is unclear what it means for an expression to be 'disjunction-like', i.e., what the property is that *at least 5* and the disjunction *5 or more* share, and why this particular property would be tied to ignorance. In effect, Büring suggests to compare *at least* to *or* in search of an explanation.

Coppock & Brochhagen (2013b) take up this suggestion. They adopt a richer semantics, unrestricted inquisitive semantics (see Ciardelli, Groenendijk & Roelofsen 2009), which enables them to define the crucial property semantically: *at least* shares with disjunctions the fact that it 'draws attention' to multiple possibilities.[19] They link this property to ignorance inferences via a maxim of 'Interactive Sincerity': don't draw attention to multiple possibilities unless you're unsure which ones are the case. A crucial feature of both Büring's and Coppock & Brochhagen's account is that the ignorance inference is linked directly to a stipulated semantic difference without taking context into account. Büring's instantiation of the Maxim of Quantity, unlike Grice's, makes no reference to a conversational goal, and Coppock & Brochhagen's

19 Their notion of 'drawing attention' is different from the notion of highlighting we employed following Coppock & Brochhagen (2013a), who invoke highlighting in addition to attention.

Interactive Sincerity likewise predicts ignorance independently of context (as long as the maxim is observed, but no proposal is included for when and how speakers might opt out of it).

The context-independence of ignorance inferences in these approaches is incompatible with our experimental results. By ignoring the role of contextual underspecification, the approaches thus far have been looking for a property of *at least* that would convey ignorance. We think that the 'mere' tendency for *at least* to be more frequently/typically used in precise contexts may be sufficient to account for the pattern of ignorance inferences across modifier and QUD types, and anything substantially stronger will not be able to account for this pattern without further stipulations.

Previous experimental work. Contrasts like the one in (6) have been found in experiments using a validity judgment task (Geurts et al. 2010), but not in experiments using a picture verification task (*In this picture, is it true that there are at least four butterflies?*), where no ignorance inferences were detected at all (Coppock & Brochhagen 2013a). One way to explain this is to conjecture that truth judgments are insensitive (or at least less sensitive) to 'mere' pragmatic inferences. However, Coppock & Brochhagen dismiss this, rightly we think, because the judgments they obtained do show other effects that seem pragmatic (in particular, an effect they claim is due to the 'highlighting' that *at most* does). The explanation Coppock & Brochhagen pursue is that there are 'weak' and 'strong' pragmatic requirements (hence inferences), and that only the strong ones may affect a truth judgment. As far as we can tell however, this division and particularly the classification of ignorance inferences as weak are not independently motivated.

A plausible explanation seems to emerge if we take into account how participants guess what the context is like. The task (to judge whether something is true) resembles the context with a polar question in (3) in which no precise answer was required, and hence no ignorance inference was predicted to occur. Since Coppock & Brochhagen (2013a) did not explicitly control for context (in the sense of QUDs), the picture verification task itself may have played the role of context, or may have lead participants to guess a context like (3). The relevant distinction to be made, then, seems to be not between strong and weak pragmatic inferences, but between those that do and those that do not require a 'precise answer' QUD. Those that arise in a precise context disappear in a contextually underspecified truth judgment task.

In contrast, a validity judgment task like Geurts et al.'s (2010) seems to be more neutral in this respect. Judging whether A implies B is (we think) like hearing A and considering whether one would in that case say B. If in such a task the context is not controlled for, participants could guess the context based on when one would typically utter A. This can explain why contrasts supposedly due to ignorance

inferences have been found in a validity judgment task. It is also why we have used a validity judgment task in our experiments, to minimize the effect of experimental task (and thereby maximize the chances of finding a contrast) when the QUD was underspecified.

Finally, experimental investigations reported in Geurts et al. (2010) and Cummins & Katsos (2010) have suggested a difference in processing cost between superlative and comparative modifiers: the former are harder to process than the latter. The results of our Experiment 1, where a contrast in reading times appears only in what we take to be an underspecified context, cast some doubt on this idea. To the extent that our self-paced reading times reflect the same cognitive processes as the various measures employed in previous work, our results suggests that the source of additional processing cost, like the source of ignorance inferences, may be primarily contextual, and any contrast found between the two types of modifiers may be due to contextual underspecification rather than the intrinsic complexity of superlative modifiers.

5 Conclusions

We have proposed, and supported with novel experimental evidence, that the ignorance inferences of numeral modifiers are primarily to be understood in purely pragmatic terms, with a difference between *at least* and *more than* coming into play only when the context is underspecified. Although contextual underspecification also plays a role in natural discourse and conversation, it is likely that the degree of underspecification present in most experimental setups is far greater. While the results obtained in such contextually underspecified experiments might be ecologically valid, they deserve closer scrutiny before they can be used as arguments for particular lexical semantic contrasts. For the same reason, such results do not necessarily support 'weaker' pragmatic theories according to which (parts of) pragmatic reasoning are fluid and defeasible. The perceived defeasibility may be in part or even wholly due to a degree of contextual underspecification that is not characteristic of naturally occurring discourse and dialogue (see Clark 1997 for a classical discussion, and Anand, Andrews & Wagers 2011 among others for a recent discussion).

References

Anand, P., C. Andrews & M. Wagers. 2011. Implicature calculation and the pragmatics of experiments. Poster at XPRAG Experimental Pragmatics Conference, Barcelona, Spain.

Baayen, R. Harald & Petar Milin. 2010. Analyzing reaction times. *International Journal of Psychological Research* 3(2). 12–28.

Barr, Dale J., Roger Levy, Christoph Scheepers & Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language* 255–278.

Bates, D., M. Maechler, B. Bolker & S. Walker. 2014. lme4: Linear mixed-effects models using eigen and s4. R package version 1.1-7. http://CRAN.R-project.org/package=lme4.

Büring, Daniel. 2008. The least at least can do. In *26th West Coast Conference on Formal Linguistics*, 114–120. Somerville, MA: Cascadilla Press.

Christensen, R. H. B. 2013. ordinal—regression models for ordinal data. R package version 2013.9-30. http://www.cran.r-project.org/package=ordinal/.

Ciardelli, Ivano, Jeroen Groenendijk & Floris Roelofsen. 2009. Attention! *Might* in inquisitive semantics. In Satoshi Ito & Ed Cormany (eds.), *Proceedings of Semantics and Linguistic Theory (SALT XIX)*, .

Clark, H. H. 1997. Dogmas of understanding. *Discourse Processes* 23. 567–598.

Coppock, Elizabeth & Thomas Brochhagen. 2013a. Diagnosing truth, interactive sincerity, and depictive sincerity. In T. Snider (ed.), *Proceedings of SALT 23*, 358–375.

Coppock, Elizabeth & Thomas Brochhagen. 2013b. Raising and resolving issues with scalar modifiers. *Semantics and Pragmatics* 6. 3:1–57.

Cummins, C. 2011. *The Interpretation and Use of Numerically Quantified Expressions*: University of Cambridge dissertation.

Cummins, Chris & Napoleon Katsos. 2010. Comparative and superlative quantifiers: Pragmatic effects of comparison type. *Journal of Semantics* 27. 271–305.

Cummins, Chris, Uli Sauerland & Stephanie Solt. 2012. Granularity and scalar implicature in numerical expressions. *Linguistics and Philosophy* 35. 135–169.

Davies, M. 2008-. The corpus of contemporary american english: 450 million words, 1990-present. http://corpus.byu.edu/coca/.

Fodor, Janet Dean. 2002. Psycholinguistics cannot escape prosogy. In B. Bel & I. Marlin (eds.), *Proceedings of the 1st International Conference on Speech Prosody*, 83–88. Aix-en-Provence, France.

Geurts, Bart, Napoleon Katsos, Chris Cummins, Jonas Moons & Leo Noordman. 2010. Scalar quantifiers: Logic, acquisition, and processing. *Language and Cognitive Processes* 25.

Geurts, Bart & Rick Nouwen. 2007. At least et al.: The semantics of scalar modifiers. *Language* 83. 533–559.

Grice, H.P. 1989. *Studies in the Way of Words*. Harvard University Press.

Jansen, C. J. M. & M. M. W. Pollmann. 2001. On round numbers: Pragmatic aspects of numerical expressions. *Journal of Quantitative Linguistics* 8. 187–201.

Kadmon, Nirit & Craige Roberts. 1986. Prosody and scope: The role of discourse structure. In *Proceedings of the Parasession on Pragmatics and Grammatical*

Theory, Chicago Linguistics Society 22nd Regional Meeting, Chicago Linguistics Society.

Krifka, M. 2009. Approximate interpretations of number words: A case for strategic communication. In E. Hinrichs & J. Nerbonne (eds.), *Theory and Evidence in Semantics*, 109–132. Stanford: CSLI Publications.

Kuznetsova, Alexandra, Per Bruun Brockhoff & Rune Haubo Bojesen Christensen. 2014. *lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package).* http://CRAN.R-project.org/package=lmerTest. R package version 2.0-6.

Nouwen, Rick. 2010. Two kinds of modified numerals. *Semantics and Pragmatics* 3. 3:1–41.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. http://www.R-project.org/.

Roberts, Craige. 1996. Information structure in discourse. In J.H. Yoon & A. Kathol (eds.), *OSU Working Papers in Linguistics*, vol. 49, 91–136. Ohio State University.

Roelofsen, Floris & Sam van Gool. 2010. Disjunctive questions, intonation, and highlighting. In Maria Aloni, Harald Bastiaanse, Tikitu de Jager & Katrin Schulz (eds.), *Logic, Language and Meaning: Selected Papers from the 17th Amsterdam Colloquium*, 384–394. Berlin: Springer.

Schwarz, Norbert. 1996. *Cognition and communication: Judgmental biases, research methods and the logic of conversation*. Hillsdale, NJ: Erlbaum.

Trueswell, John, Michael Tanenhaus & Susan Garnsey. 1994. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language* 33. 285–318.

Matthijs Westera
Institute for Logic, Language and Computation
University of Amsterdam
P.O. Box 94242, 1090 GE Amsterdam
matthijs.westera@gmail.com

Adrian Brasoveanu
Department of Linguistics
UC Santa Cruz
1156 High Street, Santa Cruz, CA
abrsvn@{gmail.com,ucsc.edu}