Word count: 3998

The role of Alternatives in the interpretation of scalars and numbers – insights from the inference task

Abstract

Numerical noun phrases (NNPs) are ambiguous between *at least two* and *exactly two* interpretations. This ambiguity has been commonly discussed as a case of Scalar Implicature (SI), where meaning is enriched by exclusion of an alternative. But the SI approach to NNPs has also been widely challenged. We tested NNPs and scalar expressions in inference tasks and found that scalars were sensitive to a manipulation that altered the relevance of alternatives, whereas *exactly* readings for NNPs were not. Our findings provide a theory-critical challenge to the SI view of NNPs and support alternative views.

Introduction

The ambiguity of noun phrases containing numerals has long been a focus of semantic/pragmatic research. Consider the following sentences with numerical noun phrases (NNPs). (1a) often carries the implication that Mary has no more than two children, whilst this implication does not always arise, as in the antecedent of the conditional in (1b). Instead, (1b) implies that if Mary has two or more children, she can claim tax relief. The interpretations of the NNP in (1a) and (1b) are often referred to as the *exactly* reading and the *at least* reading, respectively.

(1) a. Mary has two children.

b. If Mary has two children, she can claim tax relief.

In the theoretical literature, different analyses have been proposed to account for the two interpretations of NNPs. According to exhaustification views, *at least* readings are basic readings, and *exactly* readings of NNPs include also scalar implicatures (SIs), which are

derived by excluding Alternatives (Horn, 1972; Singh, 2019). SI theory explains how (2a) often carries the implication in (2c), as a result of the exclusion of the Alternative in (2b). The conditions under which a given alternative is active and/or excluded are widely thought to involve contextual relevance (Fox & Katzir, 2011; Geurts, 2010). For example, when an alternative like (2b) would address a contextually salient question, intuition suggests it would normally be excluded. Alternatively, a context for (2a) may simply require an answer to the question whether some students passed, in which case, the sense in which (2a) implies (2c) is less keenly felt.

(2) a. Some of the students passed.

b. All of the students passed.

c. Not all of the students passed.

Extending the account of 'some' to NNPs as in (1a), the Alternative, *Mary has three children,* is determined by replacing *two* with a higher value, *three*. The *exactly two* reading is then explained by conjoining the basic, *at least two* reading and the negation of the Alternative.

Many researchers argue against the Scalar Implicature (SI) view of NNPs and suggest that their interpretation should be explained by other means (Breheny, 2008; Horn, 1992). For instance, it has been proposed that the ambiguity in NNPs is explained by the optional application of a maximality operator (Buccola & Spector, 2016; Kennedy, 2015). Alternatively, it has been proposed that the basic meaning of NNPs is the upper-bounded, *exactly* reading, while the lower-bounded, *at least* reading is derived from mechanisms such as semantic (Geurts, 2006) or pragmatic (Breheny, 2008) coercion. Taken together we refer to these analyses as the non-SI view of NNPs. These views have in common the idea that *exactly* readings are not derived by excluding alternatives. Moreover, at least some of these views propose that the *exactly* reading is the primary, or biased reading. If that is correct, then

the relevance of a higher number in context should have less impact on which meaning is derived.

The SI view of NNPs suggests that they should give rise to the same experimental outcomes as other scalar terms in some standard tests. However, evidence from different paradigms shows that NNPs behave differently from scalar terms. Using a dual-task paradigm, Marty, Chemla, and Spector (2013) tested how NNPs and scalar term *some* are interpreted under cognitive load. Previous research had demonstrated that when a secondary task places participants under memory load, they derive fewer SIs (De Neys & Schaeken, 2007). Marty et al. compared the effect of memory load for sentences with *some* and NNPs. They replicated de Neys & Schaeken's finding but found that the effect of cognitive load for NNPs is the opposite of that for *some*. More recently, using a truth-value judgment task, Dieuleveut, Chemla, and Spector (2019) investigated 'primary scalar implicatures' which theorists argue are a stage in the derivation of SIs. They found evidence for primary implicatures being derived in the case of *some*, but not NNPs. Child language data has also been understood to challenge the SI view of NNPs. Children interpret NNPs with *exactly* readings more reliably than other scalar expressions (Huang et al., 2013; Hurewitz et al., 2006; Papafragou & Musolino, 2003).

In short, there is already some evidence that NNPs and scalar terms exhibit different characteristics, and this evidence runs contrary to the view that *exactly* readings of NNPs are derived via SI. However, the empirical record does not directly bear on the main points of theoretical difference, which is whether ambiguity in NNPs results from optional exclusion of Alternatives. As recognised in Marty et al. and Dieuleveut et al., the SI view of NNPs could account for the existing data by suggesting that numerals are lexically focused and triggered *exactly readings* by default, explaining the results in Dieuleveut et al. (2019). Also in that case, it could be that cancelling *exactly readings* is cognitively demanding, and this could

explain the dual-task results of Marty et al. (2013). With respect to language development, Barner et al. (2011) showed that children can derive SIs when alternatives are made explicitly available. The developmental difference in interpreting NNPs and scalars could then be attributed to differences in the accessibility of alternatives rather than any difference in their lexical semantics.

In this paper, we present experimental research which addresses this question. The paradigm we use has been developed out of insights that have been reported about the inference task. Unlike verification tasks and other less direct tasks, the inference task elicits judgements from participants about whether certain implications follow from a statement. In particular, inference task research to date has targeted implications like those in (2c). Before we outline the main features of our study, we summarise in the next section key features of previous research which has employed the inference task.

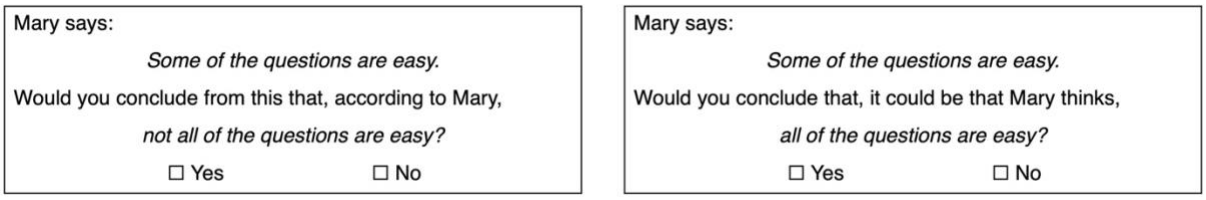

Figure 1. Examples of 'not Alt' (left) and 'could Alt' (right) conditions.

Inference Task Paradigm

Figure 1 (left) shows an example of a standard inference task (adapted from Van Tiel et al. 2016). Participants read about a de-contextualised utterance involving a scalar term and have to decide whether they could conclude that the speaker implied the negation of the Alternative. A 'Yes' response indicates that an SI is drawn. Geurts and Pouscoulous (2009) measured the rates of SIs triggered by *some* using both the inference task and a sentence-picture verification task. They found that the inference paradigm yields higher rates of SIs than the sentence-picture verification paradigm (62% vs. 34%).

The derivation of SIs we described above provides several explanations for the inflated rate in the standard inference task. One is simply that the probe in the inference task mentions the alternative, and this must raise its salience for implicature derivation (Rees & Bott, 2018). In addition, the probe question can change the context in which the sentence containing the scalar term is understood (Geurts & Pouscoulous, 2009). In particular, by asking whether you would conclude that the Alternative is false, the stimulus is suggestive that the question of whether or not the Alternative is true might be relevant to the speaker's utterance. In the absence of any other information about the context of the speaker's utterance, this may bias participants to respond 'Yes'. If this latter explanation is on the right track, we hypothesise that manipulating the contextual relevance of the Alternative via probe question should affect the rates of SIs collected from the inference task. Also, to the extent that such a manipulation is effective, it will allow us to test different approaches to NNPs. The SI view of NNPs predicts that inferences for NNPs should pattern like SIs under each probe. By contrast, the non-SI view of NNPs predicts that inferences for NNPs should be more independent from the manipulation of the probe than SIs.

The Current Study

Using a block design, we introduced two types of probe question. The 'not Alt' probe is the probe question used in the standard inference task, and a new 'could Alt' probe, which is shown in Figure 1 (right). The 'could Alt' probe asks participants whether it could be that the speaker thinks the Alternative is true. According to theoretical accounts of when scalar implicature is actually derived (mentioned above), a speaker uses *some* when *all* is true only when that more informative alternative is not relevant. Thus, in contrast to the 'not Alt' probe, the 'could Alt' probe may cue participants to this kind of context and thus bias participants to respond without SI. Note that under the 'could Alt' probe, participants who derive the SI as part of the sentence meaning should give a 'No' response. In the following,

responses associated with SI inferences are referred to as target responses. As mentioned above, target response for the 'not Alt' probe is 'Yes'.

In the current study, we tested two representative SI expressions, *some* and *possible*, and NNPs including *three*, *four* and *five*. The two SI expressions have shown near-ceiling rates of target response in previous 'not Alt' inference tasks (e.g., 89% and 93% respectively, from van Tiel et al., 2016). By contrast, NNPs have not been tested yet in the 'not Alt' paradigm. To present our predictions on response patterns, let's first consider how participants will respond to a sentence with scalar expression (like, 'some'). For either probe, we expect there to be three kinds of participant. We expect there to be participants who give target responses because they consider the SI relevant and are confident that the speaker's intended meaning includes the SI. There could also be participants who give non-target responses because they consider the SI irrelevant and are also confident about not inferring the SI as part of the intended meaning. A third group could infer both the relevant and irrelevant contexts for the SI and be uncertain about the intended meaning. How will this latter group respond in a binary judgment task? For these participants, the sentence with SI expression is ambiguous and they would not feel in a position to conclude anything about what the speaker thinks. In this kind of case, we expect participants to be more likely to give a back-off 'No' response, irrespective of whether the probe is 'not Alt' or 'could Alt'. Notice that 'No' responses are target for the 'could Alt' probe and non-target for the 'not Alt'. This means that, if probe questions provide no hint about whether the implication is intended, we expect more target responses in the 'could Alt' condition compared to the 'not Alt' condition. However, if the probe questions suggest different contexts, then this information would help to resolve ambiguities. Particularly, in the 'not Alt' condition, participants would be more likely to

accept SIs[1]. Therefore, we may even find more target responses in the 'not Alt' condition compared to the 'could Alt' condition for standard scalar terms like 'some' and 'possible'. According to the SI theory of NNPs, the ambiguity between the *at least* reading and the *exactly* reading can be resolved by the relevance of alternatives. It predicts that we should get the same response pattern for NNPs as for other scalars.

Experiment

Participants

40 adult participants (29 females) were recruited from Prolific Academic and were paid £1.4 for their participation. All participants speak English as a native language and were naïve to the purpose of the experiment. This experiment was approved by the local research ethics committee. Participants were provided with an electronic version of informed consent before taking part.

Materials and Procedure

We tested three types of scales: the quantifier scale <some, all>, the modal scale <possible, certain>, and the numerical scale. Table 1 shows examples of target items for different scales[2]. For each scale, we constructed 6 target items and 12 control items. Control items had the same structure as target items, but the responses of controls were either clearly 'Yes' or clearly 'No'.

[1] This discussion about response options does not take into account the possibility that participants derive what is referred to a 'Primary Implicature' but not the scalar implicature itself (see Dieuleveut et al for discussion). Both intuition and experimental evidence suggests that a PI-not-SI inference is accessed in a smaller number of cases, and mainly only where there are fairly clear cues to speaker ignorance about the truth of a relevant Alternative (see Dieuleveut et al.). In any case, the logic of our design allows us to set this aside as an option because a participant who only derives the PI will give the same back-off, 'no' response in both 'not Alt' and 'could Alt' trials and because any cues to speaker's knowledge about the Alternative are the same across all expression types. Thus, according to our null hypothesis where probe type has no influence on computing readings, PI-only participants will advantage target response rates in 'could Alt' trials over 'not Alt'. If the proposal that 'not Alt' probes provide a better cue to relevant context for PI-only or SI is correct, then any PI-only respondents will make this hypothesis more difficult to substantiate since they give non-target responses, despite employing Alternatives in their reading for the sentence. In light of our findings below, it makes the test for the effect of probe on response a conservative one and considerations about PI-only responders only strengthen the case that 'not Alt' probes bias a procedure for excluding alternatives.

[2] The full list of items can be found in the supplementary file.

		not Alt probe	could Alt probe
	Mary says:	Would you conclude from this that, according to Mary,	Would you conclude that, it could be that Mary thinks,
<some, all>	Some of the questions are easy.	not all of the questions are easy?	all of the questions are easy?
<possible, certain>	It is possible that the train will arrive on time.	it is not certain that the train will arrive on time?	it is certain that the train will arrive on time?
<four, five>	Four chairs are in the room.	it is not true that five chairs are in the room?	five chairs are in the room?

Table 1 Examples of target items for different scales.

We employed a within-subjects block design[3]. In one block participants responded to the 'not Alt' probe, and then in the other block they responded to the 'could Alt' probe. Each block contained 27 items, including 3 target items and 6 control items per scale. Four lists were created, each contained two blocks. Each item only appeared once in each list, in one of the two blocks. The order of the blocks was counterbalanced over the lists.

Instructions and four practice trials were provided at the beginning of each block. The order of items was randomized for each participant in each block. An unrelated study, which took approximately 5 minutes, was inserted between the 'not Alt' and the 'could Alt' block to serve as a filler task.

Results

Five participants were removed because more than 20% of their answers to control items were incorrect. Figure 2 shows the percentages of 'Yes' responses for each scale and condition by probe type. The mean accuracy of the control items was 92% (control yes: 89%, control no: 96%).

[3] A pilot study with a between-subjects design gave qualitatively similar results.

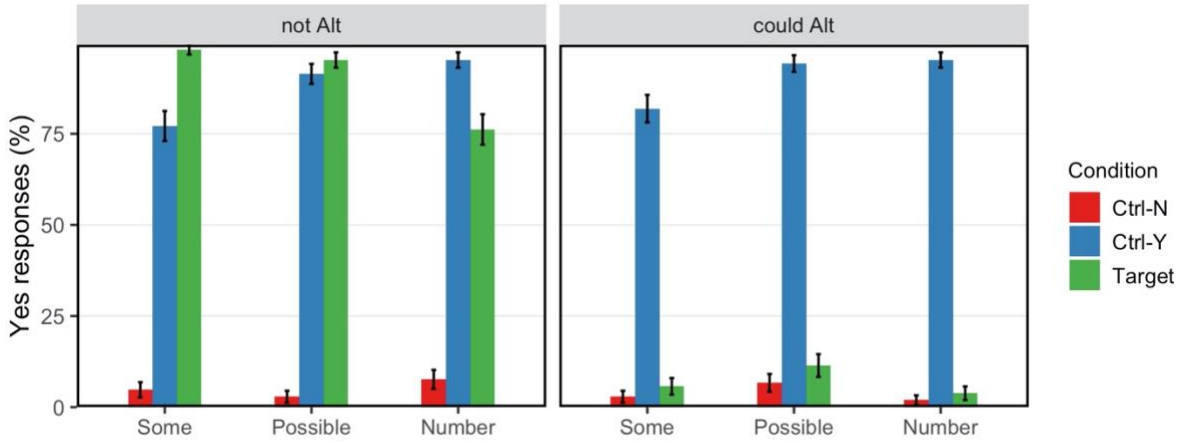

Figure 2. Percentages of 'yes' responses for each scale and condition by probe type. Error bars represent standard errors of the mean.

We coded the 'Yes' response in the 'not Alt' condition and the 'No' response in the 'could Alt' condition as target response. Figure 3 shows the percentages of target responses for each scale and probe type. We fitted a mixed effects logistic regression model predicting response (target or non-target) from probe type, scale, block order ('not Alt' first or 'could Alt' first), and their interactions, including random intercepts for participants[4]. Random slopes were dropped due to non-convergence or singularity.

[4] Mixed-effect analyses were conducted in R (R Core Team, 2020) using the 'lme4' package (Bates et al., 2015) and the 'lmerTest' package (Kuznetsova et al., 2017). Scale was dummy-coded, probe type and block order were deviation coded. Model comparisons were conducted to test the significance of fixed effects with more than two levels, using likelihood ratio tests. Significant interactions were followed up by conducting analyses on subsets of data defined by the levels of relevant factors.

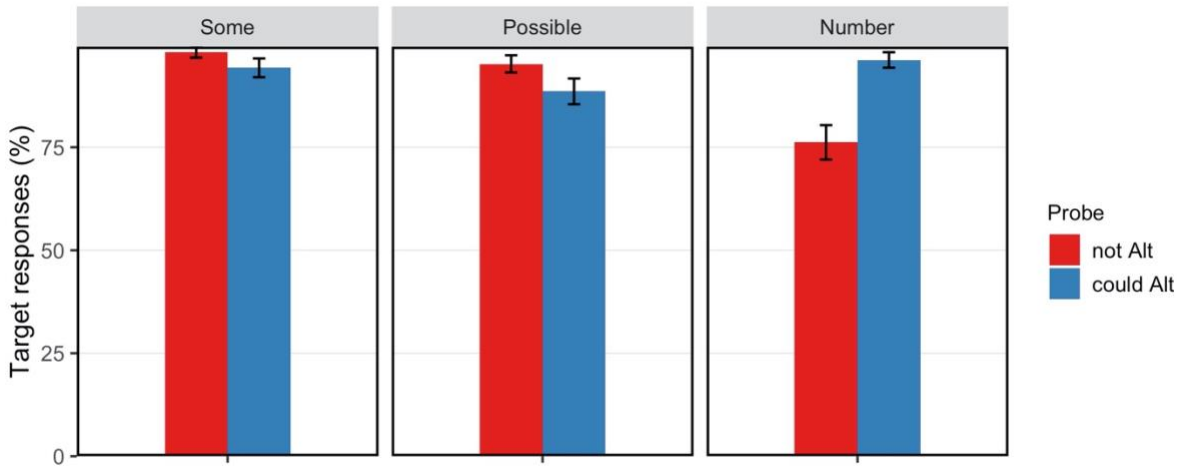

Figure 3. Percentages of target responses for each scale and probe type. Error bars represent standard errors.

There was a main effect of scale ($\chi^2(2) = 14.91$, $p < .001$) and a significant interaction between scale and probe type ($\chi^2(2) = 22.64$, $p < .001$). Planned comparisons within the levels of scale revealed that the probability of target responses was higher in the 'not Alt' condition compared to the 'could Alt' condition for 'possible' ($\beta = -1.12$, SE = 0.63, $p = .06$), the same effect was also numerically present for 'some' ($p = .16$), but it was reversed for NNPs: the probability of target responses was significantly higher in the 'could Alt' condition compared to the 'not Alt' condition ($\beta = 2.58$, SE = 0.66, $p < .001$). Planned comparisons within the levels of probe type revealed that the probability of target responses was higher for scalars than for NNPs in the 'not Alt' condition (some: $\beta = -3.17$, SE = 0.81, $p < .001$; possible: $\beta = -2.16$, SE = 0.57, $p < .001$). In contrast, in the 'could Alt' condition, the probability of target responses was higher for NNPs than for 'possible' ($\beta = 1.35$, SE = 0.64, $p = .03$) and there was no difference between NNPs and 'some' ($p = .50$).

There was also a significant three-way interaction between scale, probe type and block order ($\chi^2(2) = 8.51$, $p = .01$). Post-hoc analyses revealed that the interaction between probe

type and scale was stronger in the 'could Alt' first group ($\chi^2(2) = 17.76$, $p < .001$) than that in the 'not Alt' first group ($\chi^2(2) = 13.44$, $p = .001$)[5].

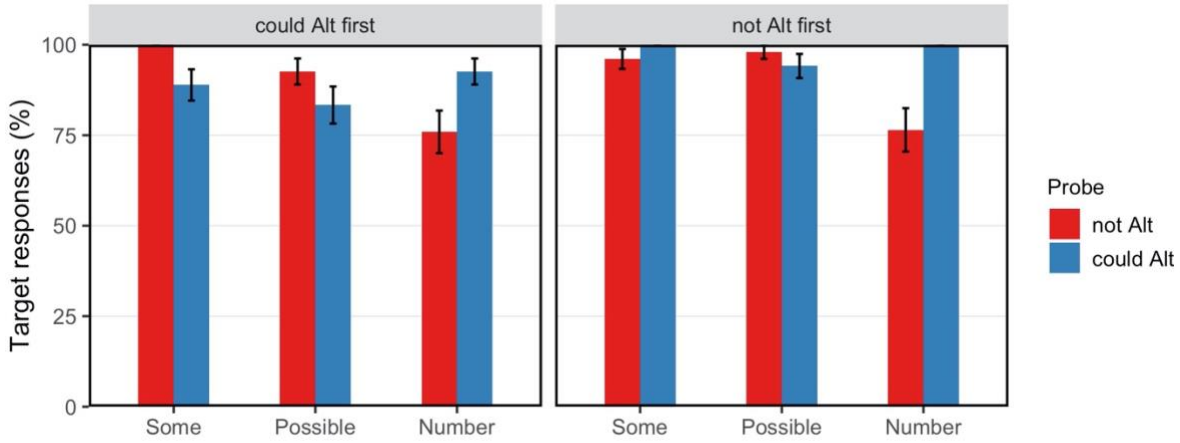

Figure 4. For each block order, the percentages of target responses for each scale and probe type. Error bars represent standard errors.

Discussion

Contrary to the predictions of the SI view, we found different outcomes for scalars and NNPs. Focusing on 'some' and 'possible', the rates of target responses were high across conditions, which may be due to the presence of alternatives ('all', 'certain') in both types of probe. Nevertheless, rates for 'not Alt' were greater than 'could Alt'. This response pattern is predicted by both views, and it reflects that context suggested by the probes resolves ambiguities in sentences with scalar expressions. Turning to NNPs, the reverse happened, target response for 'could Alt' was greater than 'not Alt'. This pattern was predicted if the probe type had no effect on ambiguity resolution. As this was the finding for NNPs, the results suggest that the relevance of putative alternatives did not help to resolve the ambiguity of NNPs. This finding is incompatible with the SI view of NNPs. However, it is in line with

[5] The interaction with block order may be due to 'not Alt' trials potentially priming participants to respond with more SIs in the following 'could Alt' block.

non-SI view that exclusion of Alternatives is not involved, or less involved, in the case of NNPs.

As this is the first inference task study to report results for NNPs and other scalars together, it is worthwhile commenting further on the patterns of responses here to highlight how the results challenge the SI view. Previous verification tasks show that *exactly* readings of NNPs are clearly more robust than SI readings for 'some' (Papafragou & Musolino, 2003; Dieuleveut et al., 2019, Huang et al., 2013, among others). An alternative account of these results suggests that NNPs do trigger scalar implicatures but do so in virtue of Alternatives being lexically focused and the SI being default (Dieuleveut et al., 2019; see also Singh, 2019). But here, in what is supposed to be the standard ('not Alt') inference task, NNPs gave rise to a clearly lower rate of target responses than scalars. If the exclusion of the NNP Alternative is default, it is difficult to see how a probe which encourages exclusion would lower the target response in this way.

Turning to the question of factors that lie behind response choice, and to elaborate our discussion above, we can draw an analogy between the ambiguity of NNPs and the lexical ambiguity like 'bank'. The sentence 'Many saw a bank' could be interpreted in its river sense or its financial sense. If participants are asked whether they could conclude that the speaker implied Mary saw a financial institution, then their responses depend on how frequent the financial sense is. Let us assume that both senses are equally frequent. Then we would expect most participants to be reluctant to say 'yes' and so they should give a back-off 'no' response. To the extent that the financial sense is more frequent, we expect a higher rate of 'yes' responses, since more participants would be likely to feel confident that the dominant sense is intended. However, if the non-dominant sense is relatively frequent, then we should still expect a group of participants to give back-off responses. Thus, any 'yes' responses to the 'bank' probe, would indicate some potential dominance for that reading; and the more

'yes' responses, the more dominant. We argue that the same reasoning applies to NNPs in the 'not Alt' condition, and this means that the *exactly* reading is quite dominant, with the *at least* reading a not insignificant minority, and this is in line with rates reported in verification tasks with numerals, such as Deiuleveut et al. Turning now to our account of why *exactly* readings for NNPs are lower than scalar in 'not Alt' trials, we explain the observed rate of non-target responses as being mostly due to back-off responses. To see this, note that, in the 'could Alt' condition, the non-target response was based on the *at least* reading only, and this rate was as low as 4%. In the 'not Alt' condition, we are assuming that the non-target response rate was the outcome of participants who only derive the *at least* reading and also those who give a back-off response due to ambiguity. Given that the *at least*-only trials are quite infrequent, 'no' responses mainly reflect the back-off response. Therefore, for NNPs, the lower rate of target responses is due to the unresolved ambiguity rather than the absence of the *exactly* reading.

Finally, our results raise a question about the status of *at least* readings of NNPs on the non-SI view and how these are robustly detected in 'not Alt' tasks. The question is why deriving an *at least* reading does not lead to a scalar implicature via normal processes of exclusion. In other words, suppose that one parse of a target sentence has an *at least* reading. We are asking why a participant who considers that parse does not also consider it valid to derive an SI using a sentence with a higher number as alternative. One answer to this question based on accounts which hold that the *exactly* reading is the single grammatically encoded meaning while the *at least* reading is a result of some subsequent coercion (semantic or pragmatic). For reasons that would need to be made clearer, it could be that only grammatically encoded meanings can enter into computations of exclusion by SI, precluding *at least* readings of NNPs from such a process.

Summary

The current study goes beyond previous findings by showing a theory-critical difference between NNPs and scalar expressions. The responses for ‘some’ and ‘possible’ are affected by the manipulation of probe in a way that is to be expected if their upper-bounded meaning is derived by exclusion of Alternatives. Responses to NNPs show a pattern that would be expected if the different probes had no effect on which reading becomes available for the participant. Our findings are consistent with the view *exactly* readings of NNPs are strongly biased but not derived as Scalar Implicature. Finally, our results substantiate previous insights about why standard inference tasks yields inflated rates for scalar expressions.

References

Barner, D., & Bachrach, A. (2010). Inference and exact numerical representation in early language development. *Cognitive Psychology*, *60*(1), 40–62. https://doi.org/10.1016/j.cogpsych.2009.06.002

Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition*, *118*(1), 87–96. https://doi.org/10.1016/j.cognition.2010.10.010

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Breheny, R. (2008). A new look at the semantics and pragmatics of numerically quantified noun phrases. *Journal of Semantics*, *25*(2), 93–139. https://doi.org/10.1093/jos/ffm016

Buccola, B., & Spector, B. (2016). Modified numerals and maximality. *Linguistics and Philosophy*, *39*(3), 151–199. https://doi.org/10.1007/s10988-016-9187-2

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

Dieuleveut, A., Chemla, E., & Spector, B. (2019). Distinctions between primary and secondary scalar implicatures. *Journal of Memory and Language*, *106*, 150–171. https://doi.org/10.1016/j.jml.2019.02.008

Fox, D., & Katzir, R. (2011). On the characterization of alternatives. *Natural Language Semantics*, *19*(1), 87–107. https://doi.org/10.1007/s11050-010-9065-3

Geurts, B. (2006). Take `Five': The meaning and use of a number word. *Non-Definiteness and Plurality*, *1*, 311–329.

Geurts, B. (2010). *Quantity implicatures*. Cambridge: Cambridge University Press. https://doi.org/10.1016/j.neuroimage.2009.02.016

Geurts, B., & Pouscoulous, N. (2009). Embedded implicatures?!? *Semantics and Pragmatics*, *2*(4), 1–34. https://doi.org/10.3765/sp.2.4

Horn, L. R. (1972). *On the Semantic Properties of Logical Operators in English.* University of California Los Angeles.

Horn, L. R. (1992). The Said and the Unsaid. *Semantics and Linguistic Theory*, *2*, 163–192. https://doi.org/10.3765/salt.v2i0.3039

Huang, Y. T., Spelke, E., & Snedeker, J. (2013). What Exactly do Numbers Mean? *Language Learning and Development*, *9*(2), 105–129. https://doi.org/10.1080/15475441.2012.658731

Hurewitz, F., Papafragou, A., Gleitman, L., & Gelman, R. (2006). Asymmetries in the Acquisition of Numbers and Quantifiers. *Language Learning and Development*, *2*(2), 77–96. https://doi.org/10.1207/s15473341lld0202_1

Kennedy, C. (2015). A "de-Fregean" semantics (and neo-Gricean pragmatics) for modified and unmodified numerals. *Semantics and Pragmatics*, *8*(10), 1–44. https://doi.org/10.3765/sp.8.10

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/h10.18637/jss.v082.i13

Marty, P., Chemla, E., & Spector, B. (2013). Interpreting numerals and scalar items under memory load. *Lingua*, *133*, 152–163. https://doi.org/10.1016/j.lingua.2013.03.006

Papafragou, A., & Musolino, J. (2003). Scalar implicatures: experiments at the semantics–pragmatics interface. *Cognition*, *86*(3), 253–282.

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Rees, A., & Bott, L. (2018). The role of alternative salience in the derivation of scalar

implicatures. *Cognition*, *176*, 1–14.

Singh, R. (2019). Context, Content, and the Occasional Costs of Implicature Computation. *Frontiers in Psychology*, *10*, 2214. https://doi.org/10.3389/fpsyg.2019.02214

Van Tiel, B., Van Miltenburg, E., Zevakhina, N., & Geurts, B. (2016). Scalar diversity. *Journal of Semantics*, *33*(1), 137–175. https://doi.org/10.1093/jos/ffu017