Review of *Embedded Scalars, Preferred Readings & Intonation: An Experimental Revisit* for Journal of Semantics

1 General

The paper presents a new experimental paradigm that aims to determine what the preferred or salient interpretation(s) of the scalar expression *some* is in order to adjudicate between the main competing semantic-pragmatic analyses. The new experimental data proves problematic both for the traditional Gricean view (and its various neo-Gricean implementations) as well for the recent grammatical approach.

The article makes a relevant and notable contribution to the scalar implicature debate, as it provide a novel and clever method at distinguishing between the different predicted readings, namely the literal, local, and global readings of *some*. In addition, the study uses auditory stimuli in order to control for (and examine) the effect of prosody on the preference and salience of the various readings. Finally, the results suggest that the available accounts need to be refined in order to account for the data.

That said, the methodology itself raises questions regarding the interpretation strategy participant may have adopted in order to complete the task at hand. What guided participants may have been independent of the actual salient and preferred readings and therefore may have biased them in a way that could have masked what the actual preferred readings were. More specifically: there is a concern that participant may have responded as soon as a reading was compatible with the state of affairs regardless of whether that reading was indeed the preferred reading.

An elaboration on the possible confounds that this new and useful paradigm may introduce is warranted and is blatantly missing in the discussion of the results of the experiment and the general discussion. The addition of such discussion will strengthen the article, sharpen the analysis of the data by ensuring that the authors understand what guided participants in the task, and will instil confidence in other scholars that the new paradigm is well understood and that response patterns that come out of using it are clear and replicable.

Reviewer decision: **Revise and resubmit**

2 Remarks on the experimental paradigm

The major contribution of this paper is the introduction and implementation of a novel experimental paradigm, the incremental verification task, that examines what the preferred reading, or readings, is of *some*. The idea is that if participants are shown a series of pictures in which each subsequent one reveals additional information that helps them decide whether a sentence is true or false, the assumption is they'll make a decision as soon the picture corresponds to an available reading. This decision will indicate which readings are available and variation in decisions and their timings will point to the preferred reading(s).

The authors convincingly argue that the incremental verification task avoids the pitfalls of previous methodologies (e.g. in Clifton and Dube 2010) that may have biased participants to accept sentences on a weaker and probably dispreferred reading or that may have led to a a strong typicality effect (Chemla and Spector 2011).

The main problem of the incremental verification task is that the stimuli are constructed in such a way as to make the literal reading the first one to be available (in step 4, namely). Suppose that participants are efficient and labour- and time-saving and therefore will likely choose to make a decision as soon as possible. That is, the first step in which participants can make a decision with certainty is step 3, which happens to correspond to the literal reading (and conveniently, doesn't rule out the validity of the local and global readings). This raises the worry that this task with respect to the *all... some* sentences is only successful at determining what the first compatible reading is that participants go with.

If in fact participants do make a decision as soon as there is any (and not necessarily the preferred) reading that matches the pictorial information, then how can the preference to answer later in the control items be explained? If participants wait till one of the readings in the structurally ambiguous sentences in the control items is false, why don't they do so in the *all... some*? One version of the grammatical analysis of scalar implicatures assumes that logical forms of sentences with scalar expressions like *some* are ambiguous between a structure with the exhaustive operator and a structure that lacks it. If this is indeed true, then the difference between the *all... some* sentences and the controls is even more puzzling. A possible explanation for the difference between the controls and critical items is that when participants are requested to choose between two readings of a structurally ambiguous sentence, they need to decide between two distinctive LFs or propositions, whereas in the case of the critical items, participants are required to distinguish between different parses, only one of which is associated with the literal reading of the sentence. And so, even if participants say YES early and it turns out that some of the parses is false, the literal reading will still be true. In the case of the target items, then, the most inclusive one is the literal one, and indeed this is what participants choose, whereas in the controls, saying YES early might prove to be wrong when all circles are revealed and the proposition that both circles and squares contain suns turns out to be true.

The hypothesis that participants go with the most inclusive reading that is compatible

with the pictorial information seems to account for the responses in the *exactly one... some*. In this condition, the first reading participants can make a decision about is the global one, which, given step 2 in Figure 5, is false. And so, the only reading the entails the other two readings is the first one out. Step 3 then rules out the literal reading. If the local reading is still available, then why didn't participants wait to see if it is indeed true as more pictorial information is revealed? Here the account the authors provide, that the literal reading is the preferred one (followed, at least in this condition, by global and then the local), is convincing: Participants waited until the preferred reading is falsified (or, in *all... some*, confirmed) to make a decision. An alternative interpretation is also possible, however: Under a surface reading of the sentence, the pictorial information necessary to evaluate *Exactly one bell is connected to some of its semicircles*, participants wanted to wait till two bells are revealed in order to know whether more than one bell is connected to some of its semicircles, independently of the fact that the global reading became false after step 2. The problem is finding a sequence of pictures that stacks the cards against such a strategy. Because if this is indeed what guided participants, then a hypothetical statement with *exactly two bells* would yield the same response pattern, only pushed back one step: step 3 would then be the one that rules out the global reading, step 4 would rule out the literal reading, and then step 5 would rule out the local reading. In sum, participant's hypothesized goal to check that *exactly n* is true may mask their preferred reading, as the pictorial step that sheds light on the former also corresponds with the literal meaning.

In conclusion, these concerns warrant a more careful discussion of how the experimental design may have biased participants and how such a bias, which may be an experimental artefact or a real discursive tendency of interlocutors, should be factored into the author's model of preferred readings. On a more general note, participants' tendency to choose a construal that doesn't rule out stronger readings is the opposite of what many analysis of conversational implicatures assume: that comprehenders would likely assume an informationally stronger, i.e. more restrictive, reading rather than a weaker one. The authors' thoughts on these surprising findings are also welcome in the general discussion.

3 Comments on the interpretation and discussion of the results

Page 39: Related to the concerns raised in the previous section, a note on the authors discussion of what the QUD is in their experiment. As they say, the QUD would be a very artificial concept here. How about the idea that given the task, the question (not phrased as an open proposition, obviously) is "give me the picture that best matches the sentences," with a bias toward the first unambiguous reading that matches the picture.

Page 41: The authors' interpretation of the data for the *exactly one* sentences seems plausible, especially given the task, in which participants are asked to match a picture

which would be described in terms of existential closure: there is a group of bells, who cardinality is exactly one, such that that group is connected with Q of its triangles.

That said, I wonder if the authors should be more cautious about incorporating the any observations from the surprising *true* responses on step 2 of *exactly one. . . some* into any semantic-pragmatic account for implicature. There's always the possibility that these cases of "quick on the trigger" responses are not guided by participants' deciding on a unique vs. non-unique interpretation of *exactly*. A similar type of early response occurred in the preference-related controls (Table 5) as well, even though an early response might turn out to be incorrect.

Page 42: The modifications the authors seems incongruous. Could the authors suggest modifications for each view that would account for both the AS and ES accounts? Maybe there is an ambiguity of uniqueness and non-uniqueness reading for numerals, modified or otherwise. If such an ambiguity story works, the authors may want to see whether it stems from a unique/non-unique ambiguity of the quantity expression *much/many* that combines with the numeral, along the lines of Nouwen 2010's S&P paper *Two kinds of modified numerals*. (See also Geurt's 2006 paper *Take five*.)