Kansas Working Papers in Linguistics, Vol. 33 (2012), 1-15

Are intermediate levels of the scale used during online comprehension of scalar implicatures?

Stephen Politzer-Ahles
University of Kansas

1. Introduction

Scalar implicature is the phenomenon whereby the use of a weak or less informative term is intended to mean that a stronger or more informative term is not true. For instance, if uttered in response to a question "Did John eat all of the cookies?", sentence (1a) will generally be understood to mean (1b).

(1) a. John ate some of the cookies.
b. =John did not eat all of the cookies.

In this example, *some* is usually interpreted as meaning *not all*. *Not all* is not part of its lexical meaning, however, as illustrated by the fact that the *not all* meaning of *some* is defeasible (can be cancelled without resulting in an illogical statement, as in (2a)), whereas the core meaning is not (2b).

(2) a. John ate some of the cookies. In fact, he ate all of them.
b. John ate some of the cookies. #In fact, he ate none of them.

Thus, the *not all* meaning must be added to the interpretation through a pragmatic enrichment process. In traditional Gricean theories of pragmatics (e.g. Horn, 1972), the interpretation of *SOME* as *NOT ALL* arises as follows.[1] *SOME* and *ALL* are both members of a lexical scale of quantifiers ordered from weakest to strongest, e.g. ⟨*SOME, MOST, ALL*⟩.[2] In this scale, *SOME* is the weakest, or least informative, term. *ALL* is the strongest, or most informative, because it entails each of the other members of the scale. Assuming that a speaker is communicating cooperatively (Grice, 1975), she will use the most informative term she can. Thus, if a speaker utters *SOME*, she must have been unable to utter *ALL*, and therefore the hearer infers that *all* was not true.

The investigation of how, when, and under what circumstances the scalar inference (the interpretation of *SOME* as meaning *NOT ALL*) arises during online processing constitutes a major area of inquiry in experimental pragmatics, and has been addressed using a wide variety of

[1] I use capitals to represent the terms *some*, *not all*, etc., as well as their equivalents in other languages. For simplicities' sake I also use these capital terms as shorthand to refer to alternative morphological forms of the same quantifier (e.g., English bare *some* and partitive *some of*). See Degen and Tanenhaus (2011) and Grodner et al. (2010), inter alia, for discussion of the influence of the morphological form of the English quantifier on scalar implicature realization; see Tsai (2004) for discussion of interpretations of different morphological realizations of *SOME* in Mandarin Chinese, the language used in the present study.

[2] It is assumed that many other classes of words fall onto lexical scales like this, e.g. ⟨*cool, cold*⟩, ⟨*like, love*⟩ (Rullman & You, 2006; Hirschberg, 1991). The scales most commonly investigated in experimental pragmatics are the quantifier scale ⟨*SOME, ALL*⟩ and the coordinator scale ⟨*OR, AND*⟩. In this study I only address the quantifier scale.

methods (speeded verification: Bott & Noveck, 2004; Bott et al., 2011; Noveck & Posada, 2003; Feeney et al., 2004; reading times: Breheny et al., 2006; Lewis & Phillips, 2011; Hartshorne & Snedeker, in press; visual world: Huang & Snedeker, 2009; Panizza et al., 2009; Grodner et al., 2009; Degen & Tanenhaus, 2011; event-related potentials: Noveck & Posada, 2003; Nieuwland et al., 2010; Politzer-Ahles et al., in press; Hunt et al., in press). These studies have reached different conclusions about whether the *NOT ALL* meaning of *SOME* is generated rapidly or at a delay, and whether or not it guides readers' predictions about upcoming lexical items as the rest of the sentence unfolds.

To the best of my knowledge, however, such investigations have focused exclusively on the processing of the *NOT ALL* meaning of *SOME*. According to the Gricean view of scalar implicature, however, a more articulated pragmatic interpretation of *SOME* should be available. The rationale for this is as follows. A speaker choosing to utter *SOME* has not only chosen not to utter *ALL*, but also spurned other alternatives such as *MOST* and *MANY*. While the semantics and pragmatics of *MANY* are relatively complicated, *MOST* is clearly a more informative term than *SOME*, as it entails *SOME* and gives more specific information about the quantity of entities being referred to.[3] For example, (3a) only tells the hearer that John ate at least one cookie, whereas (3b) tells the hearer both that John ate at least one cookie and that the quantity of cookies he ate is more than half the total available set of cookies.

(3) a. John ate some of the cookies.
b. John ate most of the cookies.

Under this interpretation, then, just as *SOME* can imply *NOT ALL*, it should also imply *not most*. This aspect of the interpretation of *SOME* was pointed out as early as De Morgan (1847:58, cited in Papafragou & Schwarz, 2005:217), who noted that "*Some* usually means a rather small fraction of the whole; a larger fraction would be expressed by *a good many*; and somewhat more than half by *most*." It is also formulated in neo-Gricean terms by Levinson (1983:133), whose definition of linguistic scales details that use of a weaker term on the scale implies negation of all stronger terms, not just negation of the endpoint:

> Given any scale of the form $\langle e_1, e_2, e_3, \ldots e_n \rangle$, if a speaker asserts $A(e_2)$, then he implicates $\sim A(e_1)$, if he asserts $A(e_3)$, then he implicates $\sim A(e_2)$ and $\sim A(e_1)$, and in general, if he asserts $A(e_n)$, then he implicates $\sim(A(e_{n-1}))$, $\sim(A(e_{n-2}))$ and so on, up to $\sim(A(e_1))$

Consider, for instance, (5a)—uttered in response to (4)—the interpretation of which is (5b) rather than (5c):

(4) Did John eat most of the cookies?

[3] There is debate regarding whether the meaning of *MOST* is composed or atomic, what the appropriate semantic representation of its meaning is (Hackl, 2009; Pietroski et al., 2009). More crucially for the purposes of the present work, there has also been disagreement over whether *MOST* is semantically or pragmatically upper-bounded (Papafragou & Schwarz, 2006; Horn, 2006; Ariel, 2004), although there is evidence in favor of the later, and several recent treatments of scalar implicature assume that *SOME* and *MOST* are on the same lexical scale (Doran et al., 2009; Katsos & Cummins, 2010; Doran et al., 2012). As the focus of the present study is on the nature of the scale used when interpreting *SOME*, rather than on the specification of the meaning of *MOST*, the reader is referred to the works cited above for further discussion.

(5) a. John ate some of the cookies.
b. John did not eat most of the cookies; he only ate some of them.
c. John ate some, and in fact did eat most, of the cookies.

Additional support for the notion that the *most* part of the scale matters for the interpretation of *SOME* is that *SOME* sounds more felicitous for describing a subset of entities that makes up less than half the set, rather than a subset that makes up more than half the set. In other words, sentence (6) sounds more acceptable if four or five out of twenty students in the class are female, and less acceptable if eighteen or nineteen out of twenty students in the class are female.

(6) Some of the students in the class are female.

Degen and Tanenhaus (2011) found that, when given a display in which a gumball machine held 13 gumballs and some number of gumballs then came out, participants were most accepting of the sentence "You got some of the gumballs" when the number of gumballs that actually came out was 5 (just under half). They became progressively less accepting of the sentence, and slower to respond, when *some* was used to refer to higher numbers of gumballs.[4] Thus, there seems to be both introspective and empirical basis for hypothesizing that *SOME* preferentially describes subsets that do not make up a majority of the set—in other words, that *SOME* might mean *NOT MOST*. Nevertheless, introspective evidence for pragmatic phenomena is not always reliable; in the present case, for example, speakers questioned directly about their intuitions about *SOME* varied considerably in whether or not they said they were willing to accept *SOME* to refer to a majority, leaving open the possibility that levels of the ⟨*ALL*, *MOST*, *SOME*⟩ scale other than *ALL* may be less salient, speakers may be implicitly but not explicitly sensitive to pragmatic bounds imposed by these intermediate levels. Thus, it is valuable to investigate this issue using an implicit rather than explicit measure, in a design that allows for

The current study investigates the nature of the scale used in the evaluation of *SOME*, specifically, whether *SOME* is interpreted not only as the negation of the endpoint of the scale (*ALL*) but also as the negation of intermediate levels like *MOST*. To investigate this, I used a speeded version of the picture-sentence verification paradigm (Wu & Tan, 2009; Tavano, 2010), contrasting participants' response times to *SOME* sentences when the entity quantified by *SOME* refers to a subset of entities that makes up the majority, minority, or precisely half of the set of entities in a preceding picture. Mandarin Chinese speakers viewed pictures that included groups of people doing different things (e.g., several people eating hot dogs and several people eating cake), and then read sentences such as the Chinese equivalent of "In the picture, some of the people are eating hot dogs". Their task was to judge whether the sentence was consistent with the picture, and to make this judgment as quickly as possible after seeing the critical word (the object that corresponded to either the majority, minority, or precisely half of the entities in the picture). The hypothesis tested was that the interpretation of *SOME* as meaning *NOT MOST* would cause response times to be faster in a context where less than half of the people in the picture are eating hot dogs than in a context where more than half the people are (for example). Such a difference could be due to predictive processing—the lower-bounded interpretation of

[4] Participants were also less accepting and slower on sentences when *some* was used to a number of gumballs less than four; this was because numbers were also used in the experiment and participants preferred to use numbers to refer to sets of gumballs that were easily countable—i.e., sets that were in the subitizing range.

SOME may lead participants to expect the word corresponding to the minority of objects and thus be slower to recognize the word corresponding to the majority—or due to verification times—participants may be faster to verify sentences that they consider more consistent with the picture, in which the word mentioned corresponds to the minority of objects.[5]

2. Methods

2.1. Participants

Data were collected from 38 native speakers of Mandarin (17 female, age range 18-42, mean 22.3) from mainland China who were members of the University of Kansas community. An additional three participants were excluded from the analysis because their accuracy was below 70% or their mean response times were more than three standard deviations above the mean response time for the entire set of participants. Many of the participants were bilingual in Standard Mandarin and a local language or dialect, but all participants reported that Standard Mandarin was the language they acquired earliest and the one they use most. All had normal or corrected-to-normal vision. All participants provided their informed consent to participate in the study and received payment, and all methods were approved by the Human Subjects Committee of Lawrence at the University of Kansas.

2.2. Materials and design

Forty-eight sentences and matching pictures were created to serve as the critical stimuli; they were adapted from Politzer-Ahles et al. (in press: Experiment 2), and the sentences were modified where necessary such that each followed the form "图片里 (In the picture), / 有的 (some of the) / <subjects> / <verb> / <object>" (slashes indicate how the sentences were divided into regions). The sentences were in Mandarin Chinese. The quantifier used in all critical sentences was 有的 (*yǒu -de*), the interpretation of which is roughly equivalent to English partitive *some of* (Xie, 2003; Tsai, 2004).[6] Four pictures were made to complement each sentence (see Figure 1).

Each picture was a "some"-type array (i.e., included actors that were doing different things and could thus be described using a *SOME* sentence, as opposed to a picture made up of actors that were all doing the same thing and would be described using an *ALL* sentence) and included six actors or items. In the picture array corresponding to the High condition, four of the actors (the majority) are interacting with the object that is ultimately mentioned in the sentence, and two are interacting with some other object. In the Mid picture array, three of the actors pictured are interacting with that object and three with another object; in the Low picture array two are

[5] Note that since the pictures included in the present study included no more than two subsets of objects, the predictions are identical whether *MOST* is represented as meaning "more than half" (a majority) or as "the largest subset" (a plurality).

[6] Although *MOST* was not explicitly included in any sentences in the experiment, it may still be relevant for participants' interpretation of *SOME*, as described in the Introduction. Mandarin Chinese has two common classifiers that correspond to English *most of*: 大部分的 (*dà bùfen –de*, "the big part of") and 大多数 (*dà duō shù,* "the big number of"). Like English *most of*, both of these can felicitously refer to pluralities and majorities, and both have defeasible upper bounds ("most… in fact, all"). Both have counterparts that refer to minorities: 少部分的 (*shǎo bùfen -de* "the small part of") and 少数 (*shǎo shù* "the small number of").

interacting with that object and four with another. In the Mismatch picture array, different proportions of actors are interacting with different objects (four-two, three-three, or two-four; 16 of each type, randomly distributed across the different sets), but none of the objects are those that are mentioned in the sentence. The reason for including six actors in each picture array was to ensure that the picture array in the Low condition would have at least two of the low-occurring item (because *SOME* might also imply more than one, picture arrays in which the low-occurring item only appears once were not used) and that the difference in number between low- and high-occurring items would be salient (in a picture array with five actors/items, the difference between two items and three items may be difficult to notice). In all conditions, the critical objects used in the pictures arrays (those that would be mentioned in sentences) were highly recognizable, being chosen as completions for comparable sentence frames by at least 86% of participants in a previously-conducted sentence completion test (Politzer-Ahles et al., in press: Experiment 2), and the words corresponding to those objects were always two characters in length. In this way I created three conditions with matching objects that represent the majority, minority, or half of the objects in the picture, and one condition with entirely mismatching objects; the sentences remained identical across all conditions, while the picture arrays varied.

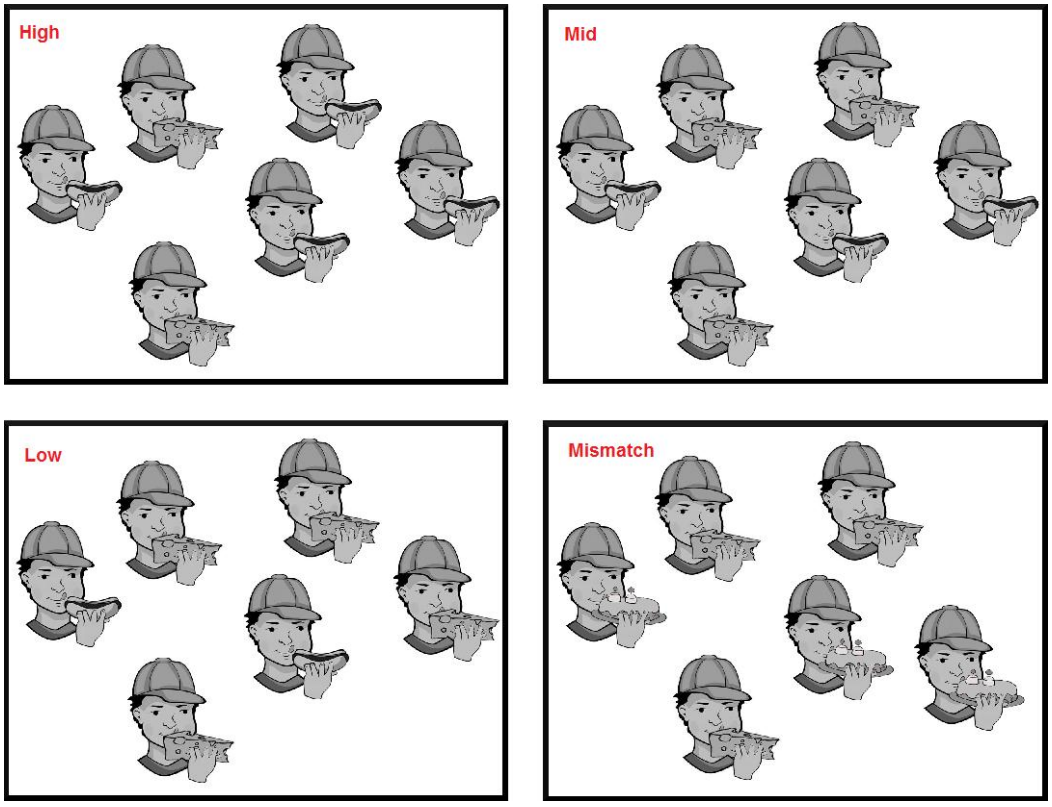

Figure 1: A sample set of picture arrays to go with the sentence 有的人在吃热狗 ("Some of the people are eating hot dogs.") The labels in the upper-left corner of each array indicate what condition each picture belonged to, and were included in the stimuli presented during the experiment. On a given trial only one of these four pictures would be presented.

The 48 experimental items were interspersed with 96 fillers of the following types: 12 "some"-type and 12 "all"-type picture arrays paired with sentences in which the verb did not match the activity described in the picture; 12 "all"-type picture arrays paired with sentences in which the object did not match the object shown in the picture array; 12 "all"-type picture arrayss paired with sentences in which the quantifier "some" did not match the number of items in the picture array (*underinformative sentences*; see Politzer-Ahles et al., in press; Hunt et al., 2011; Tavano, 2010; Nieuwland et al., 2010; Wu & Tan, 2009; Noveck & Posada, 2003); 12 "some"-type picture arrays paired with sentences with the quantifier "all", which is patently incorrect for describing these pictures (logic violations; see Politzer-Ahles et al., in press; Tavano, 2010; Wu & Tan, 2009); and 36 "all"-type picture arrays paired with sentences that correctly matched the picture. Overall the number of "some"- and "all"-type picture arrays and sentences was equal and the different picture and quantifier types were each followed by incorrect sentences 50% of the time (underinformative sentences were counted as incorrect).

The 48 experimental items were organized into four experimental lists following a Latin square design, such that each list had 12 items from each condition, no items were repeated within a list, and each item appeared exactly once in each condition and once in each list. The same 96 fillers were included in all four lists. The lists were divided into three blocks such that each block contained 4 items from each condition (or 12 items from the filler condition with sentences that correctly match "all"-type pictures). Each of the three blocks of the four lists was pseudorandomized according to the following constraints: no more than three items from the same condition occurred consecutively; no more than three correct or incorrect items occurred consecutively; no more than four "some"- or "all"-type pictures occurred consecutively; and no more than five "some"- or "all"-type sentences occurred consecutively. The four lists were pseudorandomized together, such that a given item appeared in the same order (although in different conditions) on each list.

2.3. Procedure

Participants were seated comfortably in a dimly-lit room, in front of a CRT monitor. The pictures and sentences were presented at the center of the screen using the Presentation software package (Neurobehavioral Systems, http://www.neurobs.com). Each trial began with a picture array presented in the center of the screen (at 768x576 pixels on a 1024x768px screen) for 7000 ms to ensure that subjects had time to carefully inspect the picture and fully process the relative proportions of different objects. After the picture array disappeared, it was immediately followed by a fixation point of random duration of up to 2000 ms, after which the presentation of the sentence began immediately. Sentences were presented at the center of the screen in 30pt SimSun font using the serial visual presentation procedure, with a stimulus onset asynchrony of 800 ms (words presented for 400 ms each, followed by a blank screen for 400 ms), which has been found to be natural and comfortable for Chinese readers in recent event-related potential studies (e.g. Li & Zhou, 2010). The final word of each sentence, the object, was presented simultaneously with the judgment prompts 一致 and 不一致 ("consistent" and "inconsistent") in the lower left- and right-hand parts of the screen, and participants were instructed to respond to the prompts as quickly as possible once the object word appeared. The participants' task was to judge whether or not the sentence just presented was consistent with the picture array. Because the fillers included underinformative sentences, which can be interpreted as consistent or as inconsistent with the context (Noveck & Posada, 2003; Bott & Noveck, 2004; Tavano, 2010),

participants were given explicit instructions before the experiment began that what the experimenters were interested in was the participants' own linguistic intuitions and that for some trials there would be no right or wrong answer but that they should simply respond as quickly as possible with the first judgment that came to their mind. Participants submitted their responses with their right index and middle fingers using a gamepad, and the buttons used for "consistent" and "inconsistent" responses (as well as the corresponding screen locations of the judgment prompts) were counterbalanced across participants. The trial ended with the participant's response, and the screen remained blank for 2500 ms, after which time the next trial began. Participants were given five breaks during the experiment (once every 24 trials).

Participants completed a brief practice session of 10 trials before the main experiment. The practice session consisted of items similar in structure to the sentences used in the main experiment, and included violations at the quantifier, subject, verb, and object positions. The entire experiment, including practice and breaks, took about 45 minutes.

2.4. Data analysis

Incorrect responses were removed from the analysis, as were responses faster than 200 ms or slower than 6000 ms (based on Bott & Noveck, 2004).[7] Response times were logarithmically transformed, and average response time for each of the critical conditions (High, Mid, Low, and Mismatch) was computed for each participant based on the remaining data.[8] These times were submitted to a one-way repeated measures ANOVA, with degrees of freedom adjusted using the Huynh-Feldt procedure for comparisons with more than one degree of freedom in the numerator. Additionally, the proportion of times a given participant responded "consistent" to the underinformative sentences (i.e., the proportion of logical responses) was calculated.

3. Results

3.1. Accuracy and proportion of pragmatic responses

On average, participants accepted 90% of High trials, 89.3% of Mid trials, 86.4% of Low trials, and 27.2% of Mismatch trials. After coding "consistent" responses as correct for High, Mid, and Low, and "inconsistent" as correct for Mismatch, a repeated measures ANOVA confirmed that accuracy differed across conditions, $F(1.93, 71.30) = 15.06$, $p < .001$). Bonferroni-corrected post-hoc *t*-tests showed that accuracy was lower on Mismatch trials than on all other conditions (all ps $> .001$) and none of the other conditions differed in accuracy (all $p = 1$).

Trials in which an "all"-type picture is followed by a sentence using *SOME* (for instance, a picture of six boys who are all eating hot dogs, followed by a sentence "Some of the boys are eating hot dogs"—in other words, underinformative trials) can be interpreted as either consistent

[7] In addition to these trimming criteria I also repeated the analysis using several other sets of criteria, including 0-3000 ms (Noveck & Posada, 2003), 200-3500 ms (based on the standard deviation of the current dataset), subject-specific criteria based on each subject's mean and standard deviation or inter-quartile range, and no trimming. As the pattern of results did not differ across trimming methods, only one is reported here.

[8] I also calculated means using residual reaction times (predicted reaction times for each subject were calculated using the number of characters, total number of strokes, and average strokes per character of the corresponding trial's sentence-final object). Both dependent measures were trimmed using the criteria that were applied to the raw reaction time analysis. Since the pattern of data did not differ across dependent measures, I only report the log reaction time analysis here.

or inconsistent with the context (both responses were counted as correct). A response of "inconsistent" is a pragmatic response, because the participant is evaluating the sentence based on the pragmatic interpretation of *SOME*, which is *NOT ALL*. A response of "consistent" is a logical response, because the participant is using the logical interpretation, *at least one*. A similar possibility exists for High trials: if participants interpret *SOME* as meaning *NOT MOST*, the sentence would be inconsistent with the picture. Nonetheless, the rate of logical response to underinformative trials, like High trials, was very high, 86% across participants. Unlike several previous studies that reported wide variation in participants' responses to underinformative sentence and a bimodal distribution between pragmatic and logical responders (e.g. Hunt et al., 2011; Noveck & Posada, 2003), in the present study acceptance rates to High, Mid, Low, and underinformative sentences all followed negatively-skewed unimodal distributions with high means, minima near 50%, and standard deviations near 10%, indicating that almost all participants were logical responders and that the distribution of responses to underinformative sentences was similar to that of responses to the critical sentences. Participants' acceptance rate for underinformative sentences was not significantly correlated with their acceptance rates for any other sentence types (ps > .1).

3.2. Reaction times

The comparisons of relevance to the hypotheses described in the Introduction are those between the High, Mid, and Low conditions. Comparisons between the Mismatch condition and the other conditions involve comparing across different responses (Mismatch elicits a response of "inconsistent" and the other conditions generally elicited responses of "consistent"). Nevertheless, in order to make a comparison between potential quantity-based and lexically-based effects, all four conditions were included in the statistical analysis.

The mean response times for each condition are shown in Figure 2. There difference between conditions was significant ($F_1(3,111) = 30.24$, $p < .001$; $F_2(3,138) = 31.41$, $p < .001$). However, the significant effect in the omnibus ANOVA was due to the Mismatch condition, which was responded to significantly more slowly than any other condition (all p_1s < .001; all p_2s < .001); none of the comparisons between High, Mid, and Low approached significance (all p_1s > .444; all p_2s > .681).

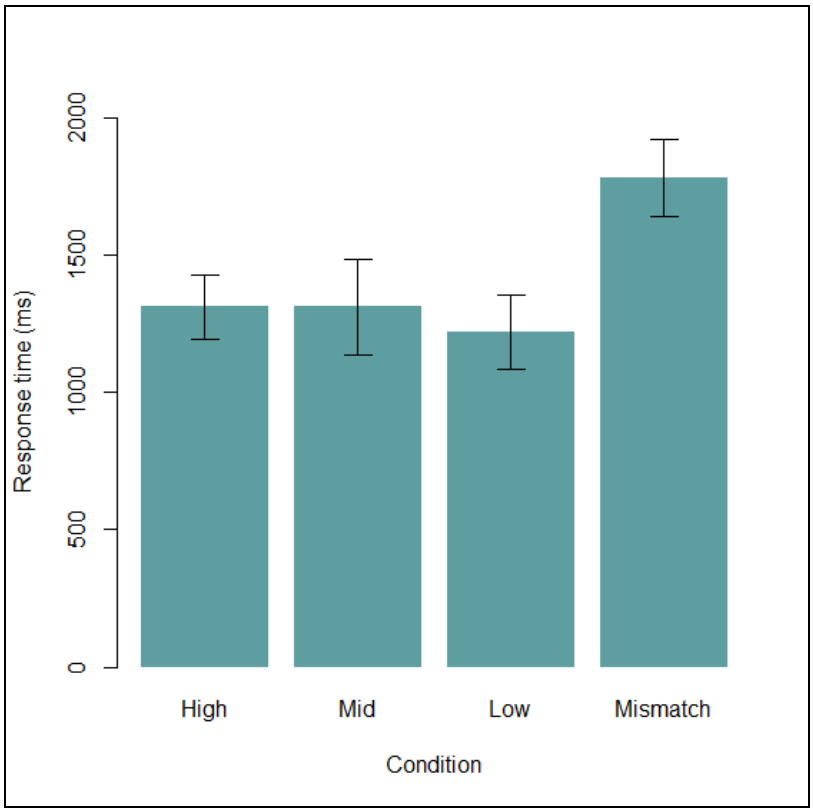

Figure 2. Mean response time for each condition.

In order to better account for potential subject-related and item-related variation in reaction times, the data were additionally analyzed using a linear mixed effects model (Baayen et al., 2008) including condition as a fixed effect and subjects and items as random effects. The results of the linear mixed effects model are shown in Table 2 (with the Low condition used as the baseline). Once again, while there was a trend towards Low having faster reaction times than High or Mid, the differences were not significant; only the Mismatch condition significantly differed from the others.

	Coefficient (log RT)	**Standard error**	***t***	***p***
Intercept	6.97	0.06		
High	0.05	0.03	1.53	.120
Mid	0.05	0.03	1.57	.130
Mismatch	0.40	0.05	7.31	<.001***

Table 1. Fixed-effect coefficients from the linear mixed effects model. *p*-values are estimated from a two-tailed *t* distribution with *df*=1522.

4. Discussion

4.1. Reaction times

The present study assumed that, if all levels of a lexical scale were used to guide processing, *SOME* would be interpreted online as meaning *not most* as well as *NOT ALL*, participants reading *SOME* sentences would thus expect the sentence to refer to an element that makes up half or less than half of a set, and subsequently they would be faster to verify sentences making reference to that element compared to sentences that are correct but make reference to a less expected element (the element that makes up the majority of objects). However, contrary to predictions, the present study found no evidence that participants were faster to verify sentences of either of the three conditions: while objects that did not match any object in the picture were responded to more slowly than objects that did, there were no differences in reaction time depending on the relative quantity of the objects in the picture.

On the surface, this finding seems contradictory to the results of Degen & Tanenhaus (2011), who demonstrated that the size of a subset influences the acceptability of and response times to sentences including *some*. It is necessary, however, to consider other reasons why response time differences may not have been observed in the present study. Alternative explanations for the null result include the absence of lexical alternatives in the materials, the task used, the salience of the difference between High and Low, and participant response strategies. These explanations are discussed below.

In the present study, the only quantifiers used in the experiment proper were 有的 ("some of") and 所有的 ("all of"). Previous empirical evidence, however, has shown that the presence of additional lexical alternatives (for example, numbers or other quantifiers) in the experimental context has an influence on how *SOME* is processed (Degen & Tanenhaus, 2011). For instance, Huang and Snedeker (2009), using a visual world paradigm with numbers included in the fillers, found that the pragmatic meaning of *some* did not rapidly influence eye movements, whereas Grodner and colleagues (2010), using a similar paradigm but without numbers in the fillers (and with several other methodological changes) found that it did. Huang and colleagues (2010) manipulated the presence or absence of lexical alternatives within a single experiment and found evidence that the pragmatic meaning of *some* was computed earlier when numbers were not included in the fillers (but not as early as in Grodner et al., 2010). Degen and Tanenhaus (2011) have interpreted these results as demonstrating that *some* is dispreferred when a more natural alternative (such as a number) is available in the experimental context. This may mean that *SOME* is only interpreted to mean the negation of another element in a scale if that element of the scale is an expected lexical alternative in the context—in other words, *SOME* can only be interpreted as *NOT ALL* if *ALL* was also available in the context, and likewise *SOME* can only be interpreted as *NOT MOST* if *MOST* was also available in the context. Support for this notion comes from both introspective and empirical data regarding boundedness (Breheny et al., 2006; Katsos & Cummins, 2010). Such data have shown that *SOME* is more likely to be interpreted as *NOT ALL* when uttered as a response to a question where *ALL* would have been a relevant response (an upper-bound context), as in (7), than to a sentence where *ALL* would not have been relevant (a lower-bound context), as in (8); examples are from Katsos & Cummins (2010).

7) a. Were all of their identity documents forgeries?
   b. Some of their identity documents were forgeries.

8) a. Is there any evidence against them?
   b. Some of their identity documents were forgeries.

Thus, if the *NOT MOST* interpretation of *SOME* is only available when *MOST* is a salient lexical alternative in the context, it would not have been available to guide processing in the present experiment. Including *MOST* in the fillers in a study like this may make this interpretation more salient.

Another possible reason for the null result is that the task may have been insensitive predictive processing. As described in the introduction, it is possible that the time taken to verify *SOME* does not differ depending on whether it refers to the minority or majority of objects, but that the interpretation of *SOME* as meaning *NOT MOST* could cause participants to make a forward expectation towards one of the objects illustrated and thus take longer to recognize the word corresponding to the other object. Differences in how strongly predicted a word is during reading are typically tested using reading times (in either self-paced reading or eye movements), electrophysiological activity (i.e., the N400 component in event-related potential research), or cross-modal priming. The present study, on the other hand, used a speeded verification task with verification responses time-locked to the presentation of the critical word. This task indeed showed a difference between response times to words that lexically matched the picture and words that did not, but this task may not be as sensitive as the other tasks described above to the subtler manipulation of pragmatic consistency. Since participants had a limited set of words they could predict (two entities in the picture that could be referred to), it is likely that they were able to predict both to some extent (see Hunt et al., 2011, for discussion of split predictions in *SOME* sentences), and the present task may not have been sensitive enough to identify subtle differences between the strengths of the two predictions. Furthermore, verification times reflect both the time needed to recognize the meaning of a word and the time taken to verify the sentence (Huang & Snedeker; but see Bott et al., 2011, for discussion of how to separate these). It remains an open question, then, whether the present study's predictions may be borne out if a different task is used.

The present study contrasted sets in which the entity referred to in the sentence constituted a High number of the entities in the picture (four out of six) versus sets in which the entity constituted a Mid (three out of six) or Low (two out of six) number of entities. Across the High, Low, and Mid conditions, the number of elements referred to always fell within the subitizing range (Degen & Tanenhaus, 2011). It is possible that the differences between these conditions were not large or salient enough to cause participants to make differential predictions. Differences in prediction strength may have been larger if the sets shown in the study were larger. The Gricean account outlined in the Introduction, however, does not predict the overall set size to matter (it only predicts differences between set sizes that correspond to different levels of a scale—such as majority, half, and minority—, not differences based on the size or salience of the difference between minority and majority). Thus, a finding of response time differences in a study with a larger difference between set sizes and not in the present study, such a result would be unpredicted by this account, and may call for a refinement of it.

Finally, another possible explanation for the null result is that participants were not paying close attention to quantification but rather were performing lexical matching between the sentential objects and the pictures. Indeed, they were far more likely to accept underinformative sentences than the participants in Politzer-Ahles et al. (in press; but see Politzer-Ahles, 2011, for a review of the variation of acceptance rates across studies and the factors that contribute to this variation). To test this possibility, I performed an exploratory analysis (using pairwise *t*-tests) of participants' accuracy in fillers which had erroneous verbs or which incorrectly used *ALL* when the quantifier *SOME* should have been used; if participants were only matching the objects to the

pictures, they would have failed to reject these types of sentences. Participants performed with 85.7% accuracy on the sentences with verb errors and 94.1% accuracy on the sentences with logical quantifier errors. (Recall that in the critical sentences, participants' accuracy was 90.1%, 89.7%, 87.7%, and 74.6% on High, Mid, Low, and Mismatch trials, respectively.) Accuracy on the sentences with quantifier errors was higher than accuracy on Mismatch ($p < .001$) and Low ($p = .005$), and marginally higher than on Mid ($p = .058$); accuracy on sentences with verb errors was higher than accuracy on Mismatch ($p < .001$) and marginally lower than accuracy on High ($p = .085$). In all, it does not seem to be the case that participants performed systematically worse in these fillers than they did on the critical trials, and in fact they were much worse at detecting semantic errors in the object position than at other positions in the sentence. This suggests that they were not merely using a lexical matching strategy in the experiment.

4.2. Acceptance rates

Participants' judgments also did not provide evidence for greater acceptance of *SOME* when referring to items that formed the minority, rather than the majority or precisely half, of a set. Rather, participants tended to accept items from all these conditions. Furthermore, there was no evidence for a split between responders who consistently accepted *SOME* referring to majorities and those who did not; rather, responders' acceptance rates formed a unimodal distribution, unlike the bimodal distributions observed in some studies on underinformative sentences (Noveck & Posada, 2003; Hunt et al., 2011). However, the same was true of the underinformative sentences included in the present study: there was not evidence for separate groups of semantic and pragmatic responders.

Picture-sentence sets similar to these were tested in Politzer-Ahles et al. (in press). When participants' task was to judge the sentences as consistent or inconsistent (Experiment 1), the majority of participants did not consistently choose one or the other, although several participants did consistently accept or reject such sentences. On the other hand, when participants' task was to rate the picture-sentence consistency on a gradient 7-point scale, two groups emerged, one of participants whose ratings showed sensitivity to the underinformativeness of the sentence (i.e., pragmatic responders) and one of participants whose ratings did not (i.e., semantic or inconsistent responders). The question, then, is why similar materials in the present study yielded mostly semantic responses.

There are suggestions in the literature that when participants are made to respond quickly, they are less likely to interpret *SOME* pragmatically (Bott & Noveck, 2004; Chevallier et al., 2008; Bott et al., 2011). Degen & Tanenhaus (2011, Experiment 2) also found high acceptance rates of underinformative English bare *some* (but not partitive *some of*) when participants were asked to respond as quickly as possible. Thus, it is possible that the nature of the task coerced most participants into behaving like semantic responders in the present study—although it should be noted that in similar verification time studies in French (Noveck & Posada, 2003; Bott & Noveck, 2004, Experiment 3) participants under time pressure still made more pragmatic responses than the present study (in terms of both overall response proportion and proportion of pragmatic responders) and exhibited grouping (in Noveck & Posada, 2003, most participants were highly consistent in their responses; Bott & Noveck, 2004, do not report response consistency for their participants, but acceptance rates to underinformative sentences do show a larger standard deviation than those to other sentence types, possibly indicative of aggregating across responders with different tendencies, and clearly different than the present

study, in which acceptance rates for underinformative sentences followed roughly the same distribution as those to correct critical sentences).

5. Conclusion

The present paper presented a previously untested prediction made by accounts of pragmatic processing that assume Gricean maxims and lexical scales are used during online comprehension—the prediction that *SOME* will be interpreted online as meaning *not most* as well as *NOT ALL*, and will guide processing accordingly. I have also presented an experimental design through which this prediction can be tested. Although the data did not provide support for this prediction—and, if proven robust, these findings would call for an explanation of why certain levels of the scale hypothesized by Gricean accounts are not used during online comprehension—several alternative explanations need to be ruled out. Chief among these alternative factors is the presence or absence of lexical alternatives in the experimental context, and the task used; both of these are fruitful avenues for future research into this aspect of scalar implicature processing and are being examined in research that is currently underway.

6. Acknowledgements

I would like to thank Jiang Xiaoming for previous discussions about this study and Robert Fiorentino for feedback on an earlier version of this manuscript. This study is based on research that was conducted at the Center for Brain and Cognitive Sciences, Peking University, through the NSF East Asia and Pacific Summer Institutes, award ID #1015160. I also thank the Linguistics Department at the University of Kansas for financial support for this experiment.

References

Ariel, M. (2004). Most. *Language, 80,* 658-706.

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390-412.

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

Bott, L., Bailey, T., & Grodner, D. (2011). Distinguishing speed from accuracy in scalar implicatures. *Journal of Memory and Language, 66,* 123-142.

Bott, L., & Noveck, I. (2004). Some utterances are underinformative: the onset and time course of scalar implicatures. *Journal of Memory and Language, 51,* 437-457.

Chevallier, C., Noveck, I., Nazir, T., Bott, L., Lanzetti, V., & Sperber, D. (2008). Making disjunctions exclusive. *The Quarterly Journal of Experimental Psychology, 61,* 1741-1760.

Degen, J., & Tanenhaus, M. (2011). Making inferences: the case of scalar implicature processing. In Carlson, L., Hölscher, C., & Shipley, T. (eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 3299-3304.

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

Doran, R., Ward, G., Larson, M., McNabb, Y., & Baker, R. (2012). A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language, 88,* 124-154.

Feeney, A., Scrafton, S., Duckworth, A., & Handley, S. (2004). The story of *some*: everyday pragmatic inference by children and adults. *Canadian Journal of Experimental Psychology, 58,* 121-132.

Grice, H. (1975). Logic and conversation. In Cole, P., & Morgan, J. (eds.), *Syntax and Semantics 3: Speech Acts*. New York: Academic press, 41-58.

Grodner, D., Klein, N., Carbary, K., & Tanenhaus, M. (2010). 'Some,' and possibly all, scalar inferences are not delayed: evidence for immediate pragmatic enrichment. *Cognition, 116,* 42-55.

Hackl, M. (2009). On the grammar and processing of proportional quantifiers: *most* versus *more than half*. *Natural Language Semantics, 17,* 63-98.

Hartshorne, J., & Snedeker, J. (in preparation). The speed of inference: Evidence against rapid use of context in calculation of scalar implicatures.

Hirschberg, J. (1991). *A theory of scalar implicature*. Ph.D. thesis, University of Pennsylvania.

Horn, L. (1972). *On the semantic properties of logical operators in English*. Ph.D. thesis, University of California, Los Angeles.

Horn, L. (2006). The border wars: A neo-Gricean perspective. In von Heusinger, K., & Turner, K. (eds.), *Where Semantics Meets Pragmatics*. Amsterdam: Elsevier. 21-48.

Huang, Y., Hahn, N., & Snedeker, J. (2010). Some inferences still take time: prosody, predictability, and the speed of scalar implicatures. *Poster presented at the 23rd Annual CUNY Conference on Human Sentence Processing.*

Huang, Y., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: insight into the semantics–pragmatics interface. *Cognitive Psychology, 58,* 376-415.

Hunt, L., Politzer-Ahles, S., Gibson, L., Minai, U., & Fiorentino, R. (in press). Pragmatic inferences modulate N400 during sentence comprehension: Evidence from picture-sentence verification. *Neuroscience Letters*.

Katsos, N., & Cummins, C. (2010). Pragmatics: from theory to experiment and back again. *Language and Linguistics Compass, 4,* 282-295.

Levinson, S. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

Lewis, S., & Phillips, C. (2011). Computing scalar implicatures is cost-free in supportive contexts. *Poster presented at the 17th Annual Conference on Architectures and Mechanisms for Language Processing.*

Li, X., & Zhou, X. (2010). Who is *ziji*? ERP responses to the Chinese reflexive pronoun during sentence processing. *Brain Research, 1331,* 96-104.

Nieuwland, M., Ditman, T., & Kuperberg, G. (2010). On the incrementality of pragmatic processing: an ERP investigation of informativeness and pragmatic abilities. *Journal of Memory and Language, 63,* 324-346.

Noveck, I., & Posada, A. (2003). Characterizing the time course of an implicature: an evoked potentials study. *Brain and Language, 85,* 203-210.

Panizza, D., Huang, Y., Chierchia, G., & Snedeker, J. (2009). Relevance of polarity for the online interpretation of scalar terms. *Proceedings of the 19th Semantics and Linguistic Theory Conference (SALT 19).*

Papafragou, A., & Schwarz, N. (2006). Most wanted. *Language Acquisition, 13,* 207-251.

Pietroski, P., Lidz, J., Hunter, T., & Halberda, J. (2009). The meaning of 'most': semantics, numerosity and psychology. *Mind and Language, 24,* 554-585.

Politzer-Ahles, S. (2011). *Online processing of scalar implicatures in Chinese as revealed by event-related potentials*. M.A. thesis, University of Kansas.

Politzer-Ahles, S., Fiorentino, R., Jiang, X., & Zhou, X. (in press). Distinct neural correlates for pragmatic and semantic meaning processing: An event-related potential investigation of scalar implicature processing using picture-sentence verification. *Brain Research.*

Rullman, H., & You, A. (2006). General number and the semantics and pragmatics of indefinite bare nouns in Mandarin Chinese. In von Heusinger, K., & Turner, K. (eds.), *Where Semantics Meets Pragmatics*. Amsterdam: Elsevier. 175-196.

Tavano, E. (2010). *The balance of scalar implicature.* Ph.D. thesis, University of Southern California.

Tsai, W.-T. (2004). Tán "yǒu rén," "yǒu de rén," hé "yǒu xiē rén" [On "yǒu rén," "yǒu de rén," and "yǒu xiē rén"]. *Hànyǔ Xuébào [Chinese Linguistics], 8(2)*, 16-25. In Chinese.

Wu, Z., & Tan, J. (2009). Scalar implicature in Chinese child language: an experimental study. *Journal of Foreign Languages, 32*, 69-75. In Chinese.

Xie, Y. (2003). Guānyú "yǒu de+VP" [On the construction of "yǒu de+VP"]. *Yǔyán Yánjiū [Studies in Language and Linguistics], 23*, 37-42. In Chinese.

Author contact information:

Stephen Politzer-Ahles: *sjpa@ku.edu*