1-1-2010

Processing scalar implicature: What can individual differences tell us?

Erin Tavano
University of Southern California, tavano@usc.edu

Elsi Kaiser
University of Southern California, elsi.kaiser@usc.edu

Processing scalar implicature: What can individual differences tell us?

Abstract

There is much current debate about processing scalar implicature, but a considerable body of empirical evidence seems to support the idea that it requires additional time and effort on the part of the hearer (e.g. Breheny, Katsos and Williams 2005, Bott and Noveck 2004 and many others). The goal of this study was to contribute to our understanding of the cognitive processes that go on as comprehenders process sentences with and without scalar implicatures. We conducted a visual-world eye-tracking experiment using a picture-verification task, with a novel single-picture display, and asked participants to indicate whether the picture they saw was a good description of the sentence they heard. As a whole, our results suggest that processing scalar implicatures does appear to entail a processing cost. In this paper, however, we take a closer look at a pattern which has also been obtained in several previous experiments (e.g. Noveck 2001, Noveck and Posada 2003), namely, the tendency for participants to split into two distinct kinds of responders in the presence of underinformative descriptions. An example of an underinformative description is "Some giraffes have long necks", which is not a sufficient description of the reality that all giraffes have long necks.

Existing research suggests that adults respond to underinformative sentences either using a consistent logical interpretation (e.g. "some" always means "some and possibly all", and thus "Some giraffes have long necks" is judged to be true) or a consistent pragmatic interpretation involving a scalar implicature (e.g. "some" always means "some but not all", and thus "Some giraffes have long necks" is judged to be false). Although it is widely assumed that participants' answers reflect their on-line processing (i.e., a logical response means that no implicature was computed, a pragmatic response means that the implicature was computed), our data suggest that participants are aware of scalar implicature regardless of how they respond to underinformative sentences, and in some cases, greater processing can be demonstrated for participants who answer "logically". We further suggest that the emergence of participant response groups may be due to participants' sense that they should be consistent within an experimental context, rather than a difference in how underinformative items are interpreted.

Processing scalar implicature: What can individual differences tell us?

Erin Tavano and Elsi Kaiser

1 Introduction

Scalar implicature refers to a phenomenon first introduced in Grice (1975), which proposed several Conversational Maxims describing rules which conversational participants obey as part of their normal efforts to be understandable and helpful. The research presented in this paper focuses on the Maxim of Quantity, which exhorts speakers to "make your contribution as informative as is required (for the current purposes of the exchange)" (Grice, 1975:45). Though this instruction seems aimed toward the speaker, in Grice's view it is also crucial for conversation that the hearer bear it in mind. If the hearer believes that the speaker has uttered something that is obviously less than fully informative, but *has* made his or her contribution as informative *as possible*, then the hearer may assume that the speaker has created an *implicature.* That is, the speaker expects the hearer to draw an inference that some obviously more informative utterance does not apply.

This idea has been formalized by Horn (1972), who introduced the idea of Horn scales. These are sets of lexical items that can be ordered such that one is regularly "above" the other in terms of informativity, and the higher item entails the lower item. Horn scales include <all, some>, <hot, warm>, <huge, big> and so on; if something is hot, then it is also warm. If a speaker says that something is "warm", then the hearer might reasonably infer that the speaker has intentionally avoided the higher term "hot", and that the item is therefore not hot, or at least the speaker does not have evidence to say that it is. In this situation, the hearer has processed the speaker's scalar implicature, from the scale <hot, warm>.

There is much current debate about processing scalar implicature, but a considerable body of empirical evidence seems to support the idea that it requires some time and effort on the part of the hearer (e.g. Breheny, Katsos and Williams, 2005; Bott and Noveck, 2004; but see Grodner, Klein, Carbary and Tanenhaus, 2008 for a different view). To further investigate the time course of processing scalar implicature, we conducted a visual-world eye-tracking experiment using a picture-verification task (see Tavano & Kaiser, to appear for details). By using a single-picture display and asking participants to indicate whether the picture they saw was a good description of the sentence they heard, we aimed to contribute to our understanding of the cognitive processes that go on as comprehenders process sentences with and without scalar implicatures. As a whole, our results suggest that processing scalar implicatures does appear to entail a processing cost, at least when involving visual contexts of the kind that we looked at.

In the present paper, we take a closer look at a particular aspect of our findings, namely the tendency for participants to split into two distinct kinds of responders, a pattern which has also been obtained in several previous experiments. More specifically, a number of researchers have noted that adults often split in how they respond to underinformative sentences (Noveck, 2001; Noveck and Posada, 2003; Foppolo, 2007). An underinformative sentence is one that uses the lower lexical item on a Horn scale where the higher one would be more appropriate. One example from Noveck (2001) is "Some giraffes have long necks." In reality, *all* giraffes have long necks, so the use of "some" is underinformative.

Existing research suggests that adults respond to underinformative sentences either using a consistent logical interpretation (e.g. "some" always means "some and possibly all", and thus "Some giraffes have long necks" is judged to be correct) or a consistent pragmatic interpretation involving a scalar implicature (e.g. "some" always means "some but not all", and thus "Some giraffes have long necks" is judged to be incorrect). Note that the pragmatic interpretation is the one which requires the scalar implicature to be computed (i.e., the speaker said "some" and thus meant "not all"). In addition, researchers have often observed a difference in response times between the two interpretations, with the pragmatic interpretation being slower than the logical interpretation. The slowdown associated with the pragmatic interpretation is standardly analyzed as indicating increased processing cost associated with scalar implicature (e.g. Noveck, 2001; Noveck and Posada, 2003). On the whole, previous research has assumed that participants' answers reflect their on-line processing (i.e., a logical response means that no implicature was computed, a prag-

matic response means that the implicature was computed). However, previous authors have had little to say as to *why* some participants seem to consistently opt for the logical interpretation while others consistently opt for the pragmatic interpretation. Additionally, previous research did not directly investigate whether participants' off-line pragmatic vs. logical responses provided a reliable indication of whether they were computing the implicatures. In particular, one might ask whether participants who provide logical answers were actually failing to compute the implicature, or whether they computed it but nevertheless opted to respond in a logical manner.

Our data suggest that participants are aware of scalar implicature regardless of how they respond to underinformative sentences, and in some cases, greater processing can be demonstrated for participants who answer "logically". We suggest that the emergence of participant response groups may be due to participants' sense that they should be consistent within an experimental context, rather than a difference in how underinformative items are interpreted.

2 Previous Experiments

The present study joins others in using the visual-world eyetracking paradigm that investigate the real-time processing of scalar implicature (Storto and Tanenhaus, 2005; Huang and Snedeker, 2009; Grodner, Klein, Carbary and Tanenhaus, 2008). With the exception of Grodner et al., the results of the other eye-tracking studies suggest that scalar implicatures induce an increased processing cost. Because our main focus in this paper is on the split into pragmatic and logical responders, and because these studies did not report any differences in participant consistency, we will not focus on their results here (for a review, see Tavano and Kaiser, to appear). However, it is important to note that the participants in these experiments all ultimately responded according to an interpretation that incorporated the scalar implicature, as was required by the nature of the experiment.

Other studies — such as Noveck (2001) — found that participants tend to split into pragmatic responders and logical responders. Noveck (2001) tested children's and adults' interpretation of "might" and "some". With "might", the logical interpretation is "might be and possibly must be", and the pragmatic interpretation is "might be but does not have to be". With "some," the logical interpretation is "some and possibly all", and the pragmatic interpretation is "some but not all". Noveck's results show that children were far more likely to give logical responses to underinformative scenarios containing "might" or "some", than adults, who tended strongly towards pragmatic responses. When adults were faced with an underinformative condition, e.g., having to accept/reject a statement such as "there might be a parrot in the box" in a situation where there must be (Noveck's Experiment 1), Noveck found that adults accepted such a sentence on only 35% of the trials. In other words, only 35% of adult responses agreed with the logical "might be and possibly must be" interpretation. In contrast, children showed a stronger preference for the logical response, with the strength of the preference modulated by the age group.

Noveck (2001) obtained a similar pattern of results with a slightly different task (requiring a truth judgment on sentences like "Some giraffes have long necks", his Experiment 3). Adults again showed an overall preference for pragmatic responses; only 41% of adult responses agreed with the logical "some and possibly all" interpretation. As before, children showed a stronger preference for logical responses. Noveck attributes these results to adults' overall greater processing ability, suggesting that it allows adults, but generally not children, to compute the scalar implicatures that are needed to generate the pragmatic response.

However, Noveck's adults were not uniform in their response patterns. In Experiment 1, 6 of 19 (32%) consistently responded logically and 13 of 19 (68%) provided consistently pragmatic responses. The division in Experiment 3 was not as strong; 5 of 15 (33%) consistently responded logically and 6 of 15 (40%) pragmatically, with the remaining 4 equivocating between pragmatic and logical responses. In light of Noveck's claim that the pragmatic responses are produced due to greater adult processing capacity, these split results raise the question of whether the logical-response subset of adults simply has insufficient, child-like processing capacity. In a follow-up to this study, Noveck and Posada (2003) suggest an explanation in terms of Relevance Theory: it may be the case that not all adults have the same threshold for deciding what input is likely to be informative enough to be worth their processing efforts. Under this view, the logical-response

adults are not necessarily child-like, but possibly just a bit stingier with their processing effort.

Noveck and Posada (2003) conducted an ERP study on implicature processing, evaluating both response time and the N400 brainwave response. The N400 wave indicates semantically anomalous stimuli; a larger N400 response indicates greater perceived anomaly. Participants were asked to give truth judgments on several types of sentences, including obviously true sentences, obviously false sentences, and underinformative sentences like "Some elephants have trunks" (underinformative because, of course, all do).

Noveck and Posada again demonstrated a split among their participants with regard to the interpretation of "some". Similar to Noveck (2001), there were more pragmatic responders than logical responders: Twelve of 19 participants (63%) gave a consistent pragmatic interpretation, (answering "false" to "Some elephants have trunks"), while the remaining 7 (37%) gave a consistent logical interpretation of "some" (answering "true", since some elephants do have trunks). The pragmatic interpretation group was also significantly slower (1064ms) than the logical interpretation group (647ms), and the delay was attributed to the time taken to process the scalar implicature.

To summarize the ERP results, Noveck and Posada looked at the N400 response to the last word of each sentence, and found that underinformative items yielded a lower response relative to the obviously true and false items, indicating that the underinformative items were not perceived to be highly semantically anomalous, at least not in the time window under consideration. Combined with the longer response time, the authors suggest that the results were due to a pragmatic process (or other decision-making process) that took place after the last word had been heard. However, there was no ERP difference between the participant interpretation groups, as one might have expected.

Another study suggesting that adults do not always provide implicature-derived answers is Foppolo (2007). Foppolo's main focus is actually the idea that processing differences occur between different types of sentence structures, rather than different types of participants. Her experiment, a picture-verification task with response time measures, used the logical connectives <and, or>. The logical interpretation of "or" is *inclusive or*, that is, saying "A or B" means "A or B and no information about whether A and B". On the pragmatic interpretation of "or" (referred to as *exclusive or)*, saying "A or B" means "A or B but not A and B". The pragmatic/exclusive *or* is thought to be derivable from the scalar implicature inherent in saying "or" rather than the stronger "and".

The experiment in Foppolo (2007) was designed to test whether scalar implicatures are costly to process in certain semantic contexts, e.g. Downward Entailing (DE). Indeed, she found that when participants read a sentence in the DE context condition, they only accepted the pragmatic (exclusive "or") interpretation 57% of the time. They were also significantly slower to accept it than reject it. When the sentence was in the non-DE context (easy) condition, there was a relatively higher acceptance (87%) of the pragmatic (exclusive "or") interpretation, and participants were slightly but not significantly slower to accept it than reject it.

For our purposes, what is relevant about these results is that neither condition resulted in full acceptance of the pragmatic interpretation. As in previous studies, we see that adults do sometimes respond logically (i.e., as if they had not generated the implicature). Although Foppolo does not report the consistency of her participants, these results at least suggest that adults do not inevitably provide implicature-requiring answers, though (again as in the previous studies) the majority do.

In sum, existing work suggests that, (i) on the whole, adults tend to have a preference for pragmatic interpretations, that is, those that reflect processing of scalar implicature, and (ii) adults tend to have a consistent preference for either pragmatic interpretations or logical interpretations. The reasons underlying the grouped or split behavior are not yet fully understood. In particular, one might ask whether the split results are due to individuals having varying amounts of processing capacity, with lower-capacity comprehenders being less likely to engage in scalar implicature processing. On a related note, it could be the case that comprehenders vary in their threshold for Relevance, i.e., the point at which they decide that certain input is worthy of processing effort (Sperber and Wilson, 1995), as suggested by Noveck and Posada (2003). It is important to note that both of these ideas rely on the assumption that comprehenders' logical vs. pragmatic responses provide an accurate reflection of whether or not the scalar implicature was computed. However, as discussed in the following sections, the results from our experiment suggest that this is not necessarily true. As a whole, our findings indicate that participants tend to be aware of the

implicature regardless of their off-line responses, and suggest that the existence of participant response groups should not be taken as evidence for fundamental processing differences.

3 Experiment

The present experiment used a new type of eyetracking experimental design for investigating the processing of scalar implicature. Existing visual-world eyetracking studies presented participants with multiple mini-scenes simultaneously which depicted pragmatic and logical interpretations of a sentence with a scalar term. By analyzing participants' eye-movements to the different scenes, one can gain insights into which interpretations are being considered.

However, in our experiment, participants were only shown one visual scene at a time, and were asked to perform a picture-verification task. With this kind of design, we did not expect looks to any particular scene or object at any particular time. Instead, we wanted to test whether participants' overall manner of observing the scene is distinct for underinformative vs. appropriately informative scenarios. In addition, participants were able to accept or reject descriptions of appropriate and underinformative scenarios, providing us with the opportunity to analyze eye movements relative to response types.

The experiment had a 2x2 within-subjects design. We manipulated quantifier type ("Some" or "All") and picture type (Match or NoMatch). To create pictures that either matched or mismatched the experimental sentences, we used two kinds of pictures, as shown in Figure 1. So-called "Picture-All" (PA) displays contain objects that are all the same color (e.g. the top and bottom pictures in Figure 1). In contrast, "Picture-Some" (PS) displays contained two groups of objects that differ in color, as shown by the middle two pictures in Figure 1. In all pictures, the objects depicted are multiple identical tokens of the same object (e.g. all apples). The two picture types were crossed with quantifier type (Quantifier-Some vs. Quantifier-All, abbreviated QS and QA), yielding the four conditions in Figure 1. Because we used a picture-verification paradigm, a participant only saw *one* of these pictures on a given trial. These four conditions can also be thought of in terms of whether or not the picture and the sentence match. We use the term "Picture Match" for conditions where the picture agrees with the quantifier (QAPA, QSPS conditions), and the term "Picture NoMatch" for conditions where the picture and the quantifier fail to agree (QAPS, QSPA).

3.1 Methodology

Twenty-four students from the University of Southern California participated in the experiment and received $10. We recorded eye movements using an SR Research Eyelink II head-mounted eyetracker sampling at 500 Hz.

The experiment used 20 target items, which contained 5 or 7 occurrences of the same object. In the Picture-All conditions, all objects in the picture were identical. In the Picture-Some conditions, the objects were identical except for color. As shown in Figure 1, in the Picture-Some conditions, the objects were presented in contiguous groups of two different colors. In these kinds of pictures, there were always 3 "target" items (those that had the color named in the sentence) while the other 2 or 4 objects were a different color. This allowed us to avoid biasing participants towards minority or majority choices. The target items were the majority in the 5-object pictures and the minority in the 7-object pictures. The 5-object pictures and 7-object pictures were seen in equal numbers by each participant. Images were counterbalanced such that target and non-target items appeared in each of the 8 potential locations in the layout (two at top, two at bottom, two at left, two at right) an equal number of times. The colors of the objects were unambiguous and natural for the object (e.g., apples were red and green, not purple.) All pictures were presented in color. The darker apples in Figure 1 were actually red and the lighter-gray apples were green.

Each trial began with a sentence naming the objects in the display: "This is a picture of apples." The second sentence was the critical target sentence, and varied across conditions as shown in Figure 1. The quantifier was not stressed.

The experiment also included 40 filler items, none of which were underinformative, for a total of 60 trials per participant.
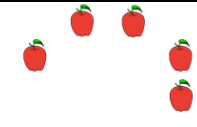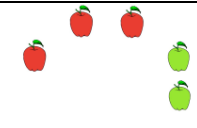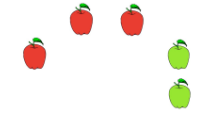
Condition	Example sentences	Example scene	Expected answer
Quantifier-All-Picture-All (QAPA/Match)	This is a picture of apples. All of them are red.		Yes (good description)
Quantifier-All-Picture-Some (QAPS/NoMatch)	This is a picture of apples. All of them are red.		No (not good description)
Quantifier-Some-Picture-Some (QSPS/Match)	This is a picture of apples. Some of them are red.		Yes (good description)
Quantifier-Some-Picture-All (QSPA/NoMatch)	This is a picture of apples. Some of them are red.		?

Figure 1: Conditions and sample stimuli

3.2 Procedure

Participants' eye movements were recorded as they looked at pictures and listened to the sentences while engaged in a picture verification task. Participants were asked to indicate whether the sentences they heard were a "good description" of the corresponding picture, and responded "yes" (good description) or "no" (not a good description) by pressing buttons on an Eyelink input unit. Before the start of the experiment, participants were told that if the sentences seemed wrong, misleading, or did not give enough information, they might be considered bad descriptions. However, participants were encouraged to use their own best judgment.

4 Results

In this section, we start by presenting information about participants' yes/no responses and response times, and then discuss the eye-movement data. In this paper, we focus specifically on (i) whether participants' off-line responses legitimately indicate a split into logical interpretation responders and pragmatic interpretation responders, and (ii) whether/how such a split is reflected in response times, eye-movement patterns and general awareness of the implicature. Please see Tavano & Kaiser (to appear) for a more detailed discussion of the general results of this experiment.

4.1 Participant responses

Condition	%"Yes" responses
QAPA	100%
QAPS	0%
QSPA	56%
QSPS	97%

Table 1: Percentage of participant responses of "Yes" ("Good description"), by condition

Overall, participants' yes/no responses show very clear patterns in three out of the four conditions. When the quantifier and picture match (QAPA and QSPS), we find an overwhelming number of "yes" answers (see Table 1). The quantifier and picture do not match in the other two conditions,

QAPS and QSPA. In the QAPS condition ("Some of them are red", all are red in the picture), we find a clear majority of "no" responses, as expected. However, in the QSPA condition, where items were underinformative, the results were mixed. We expected that underinformative sentences would not be considered to be a good description of a picture, especially since participants had been told that "not enough information" could be an example of a bad description. However, there were numerically more logical-interpretation "Yes" responses (67 of 119 responses, 56%) than pragmatic-interpretation "No" responses (52 of 119 responses, 43%) to these items. This contrasts with responses in previous experiments, which most often reflected a preference for the pragmatic interpretation in adults.

When we take a closer look at the source of these mixed responses, it becomes clear that they stem primarily from a split among our participants, rather than mixed behavior within individual participants. In fact, 22 out of 24 participants showed consistent behavior and gave the same answer on 4 or 5 out of the 5 QSPA trials. Only 2 participants out of 24 were more variable. When we look more closely within the 22 consistent participants, we find that 9 participants are pragmatic responders (consistently answering "No"), and 14 are logical responders (consistently answering "Yes").

While previous experimenters have interpreted logical responses as evidence that some participants did not process the implicature, it appears that these responses alone are not a good indicator of this. Nearly all of our participants (22 of 24, 92%) expressed awareness of the implicature, either through the experiment responses (giving the pragmatic "No" response in the QSPA condition), or in debriefing after the experiment, or both. Participants who answered logically often volunteered this information as part of an explanation for how they answered. We discuss the implications of these findings in more depth in Section 5.

4.2 Response times

In this section, we first consider participants' overall response times, and then take a closer look at the response times in the underinformative QSPA condition.

Condition	Overall RT (ms)
QAPA	1165
QAPS	1109
QSPA	1527
QSPS	1277

Table 2: Overall response times by condition.

Table 2 shows response times collapsed across yes/no responses. (We measured response times from the end of the second sentence ("Some of them are red") to the time when the participant pressed the button to indicate their response.) Response times are slowest when the quantifier is "Some". The slowest overall RT is in the QSPA condition, followed by the QSPS condition. The difference between the QSPA and QSPS conditions is significant ($p < .05$). In addition, comparing conditions where the picture is the same and the only difference is the quantifier, each "Some" condition is slower than the corresponding "All" condition (QSPA is slower than QAPA, QSPS is slower than QAPS; p's $< .05$).
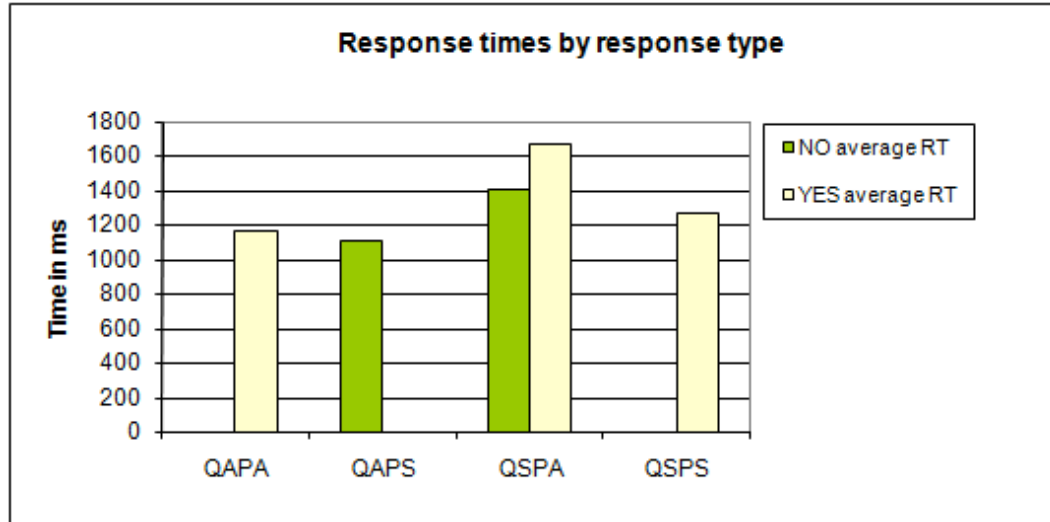
Figure 2: Response times for "yes" and "no" responses in the four conditions[1]

As mentioned in the previous section, the QSPA condition was the only one where participants were split in their responses. To better understand the nature of the slowdown in the QSPA condition, let us take a closer look at the response times for "yes" answers and "no" answers, represented by the light and dark bars in Figure 2 respectively. Figure 2 shows response times for all "yes" responses and the response times for all "no" responses, regardless of whether they were produced by a logical responder or a pragmatic responder. Looking at QSPA response times solely by response, we see that there is a difference between Yes and No responses: Yes/logical responses (1414 ms) were faster than No/pragmatic responses (1673 ms). An independent samples t-test confirms that this difference is significant ($p < .05$).[2]

Interestingly, when we look at QSPA response time for participants grouped by their overall response tendency (including the response times for the one out-of-category response, for participants who were "4 out of 5" consistent), we find that average response times are almost exactly equal: YES/logical group: 1519 ms; NO/pragmatic group 1502 ms. Thus, this suggests that the slow-down for pragmatic responses takes place on the level of each individual response rather than on the level of pragmatic vs. logical responders.

4.3 Eye movement results

We hoped to gain insight into the real-time processing of scalar implicature — in particular, to better understand the split into pragmatic responders and logical responders — by analyzing participants' eye movements. In particular, we wanted to see whether participants' overall manner of observing the scene is distinct for underinformative and appropriately informative situations.

As reported in Tavano and Kaiser (to appear), we computed the average inspection duration for each condition. An inspection was defined as the amount of time that participants spent looking at each object, even if there were multiple fixations within its bounds. We looked at inspection durations in a time frame from the second half of the quantifier, where participants should have first been able to recognize the word, to the end of the trial. The average inspection duration was longer in Quantifier-Some conditions than in Quantifier-All conditions (significant by participants, $p < .05$, marginal by items, $p < .06$). Another kind of eye movement measure (switch latency) con-

---

[1]Note that, as shown in Table 1, there were a small number of "Yes" responses in the QAPS condition (0.8%) and a small number of "No" responses in the QSPS condition (3.3%). The response times of those rare responses are not included in Figure 2 as they are presumed to be errors.

[2]However, the yes/no responses in the QSPA condition are not entirely independent, since, as previously mentioned, it occasionally happened that a person responded 4/5 times one way and 1/5 times the other. Since if we excluded these cases, there was not enough data to perform either an independent-sample or paired-sample t-test, this result should be regarded with some caution.

firms participants' sensitivity to the distinction between "some" and "all": In the Picture-Some conditions (when the picture depicted two color-groups of objects), participants were faster to look at the "other" group (defined relative to where they were looking at quantifier onset) when the quantifier was "some" than when it was "all" ($p<.05$).

Recall that the underinformative QSPA condition was the only one in which participants' responses showed a split into "yes" and "no" answers; other conditions showed an overwhelming preference for either "yes" or "no". To obtain another measure of processing load in this condition, we conducted response-contingent analyses, where we looked at the number of inspections by the type of response. We found that in the 200-600 ms period after the end (offset) of the quantifier, participants in the Yes/logical group in the QSPA condition had significantly more inspections than participants in the No/pragmatic group (3.8 inspections vs. 2.4 inspections). Furthermore, inspections initiated in this time period led to an overall significantly longer summed inspection time for the Yes/logical group (2141 ms) than the No/pragmatic group (926 ms), though there was no significant difference in average inspection times. There were no differences in inspection count or inspection time when longer time periods in the QSPA condition were considered.

5 Discussion

Our experiment aimed to provide new data on how scalar implicature is processed, by means of an eyetracking experiment with a novel type of visual scene design. In this paper we particularly focus on a pattern that cropped up both in our study and a number of previous experiments, namely the observation that participants tend to split into two distinct kinds of responders, what we have termed *pragmatic responders* and *logical responders*. Given an underinformative utterance, a subset of participants (the *pragmatic responders*) consistently performed the task in a way that seemingly reflected the processing of a scalar implicature. The remainder of participants (the *logical responders*) consistently responded as if they had not processed a scalar implicature. We define a consistent responder as one who gave the same type of response in 4 or 5 out of 5 trials in a condition. By this definition, only 2 out of the 24 participants in our experiment were inconsistent in their responses to underinformative sentences.

The questions that presently interest us are as follows: (i) Why should this response grouping occur? (ii) What can we conclude from it with regard to participants' implicature processing? Specifically, do participants' ultimate pragmatic vs. logical responses provide a reliable indication of whether or not they were computing the implicatures?

Generally speaking, it seems that previous research often assumed that participants' answers reflect their on-line processing (i.e., a logical response means that no implicature was computed, a pragmatic response means that the implicature was computed). However, our results suggest that participants' off-line responses to an underinformative utterance may not be a reliable indicator of whether they have processed a scalar implicature for that utterance. This is because we found that nearly all of our participants, both logical and pragmatic responders, eventually expressed awareness of the scalar implicature where "some" can mean "some but not all".

An interesting issue that merits further research is the question of when logical responders achieve this awareness. Our current data do not allow us to determine when this took place. For example, one might suggest that the logical responders did not process the implicature for early trials in the QSPA condition, and only thought of it during later trials or even during the debriefing. While this is possible, we think it is rather unlikely for two main reasons. First, response times for first QSPA trials tended to be longer than for subsequent QSPA trials, for all responders. If generating an implicature results in a slowdown, as many researchers have suggested, then a slowdown on early trials does not fit with the idea that participants were failing to generate implicatures specifically on those early trials. Second, there are other indications of greater processing for the logical group (more inspections, longer inspection time) relative to the pragmatic group. Therefore, we suggest that logical and pragmatic responses do not clearly indicate whether participants actually processed an implicature or not.

That said, however, our results also agreed with prior experiments in that response times for pragmatic responses were longer than those for logical responses. This result has commonly been taken as a sign that there is greater processing when scalar implicature is present. If we are claiming that responses are not an indicator of implicature processing, what explains the longer prag-

matic response times? We do not have a complete answer to this question, but feel it is important to point out that pragmatic response times are only longer when data is examined by response only, without regard to the type of responder. The logical responders and pragmatic responders actually had equal response times when examined by responder type (see Section 4.2). This suggests strongly that the division between logical and pragmatic responders is questionable. We refer again to Noveck and Posada's (2003) suggestion that the existence of responder groups can be explained by Relevance Theory (Sperber and Wilson, 1995); the threshold for relevance (a.k.a. the cost for processing the implicature, relative to its informative value) was lower for some people than others. However, it may be that the threshold is not so firmly set for any individual participant, and may actually change over the course of the experiment.

Let us return to the initial question of *why* participant response grouping should occur. We would like to suggest that — at least in our experiment — the grouping did not occur in the way that has been previously assumed (i.e., depending on whether or not the implicature is generated).

It is worth pointing out another relevant difference between the present study and previous ones: in earlier studies, adult participants most often responded pragmatically, whereas in our study, more participants were logical responders than pragmatic responders (14 vs. 9). In light of the original supposition that adults are usually pragmatic responders because of their greater processing capacity, the high proportion of logical responders seems very unexpected. In fact, there is no reason to believe that our participants had anything less than typical adult competence. If anything, our participants, nearly all university students, including several graduate students, were likely to have high inference and language skills. We tentatively hypothesize that *all* our participants processed a scalar implicature in early QSPA trials, and that they subsequently made a decision about what type of response the experimenter expected, whether an underinformative description was a "good" one or not. The equal response times between responder groups (see Section 4.2) could thus be explained by the equal time it would have taken participants to recognize an underinformative scenario and recall earlier their decision about how to respond to uninformative descriptions. It is possible that the logical responders did not do this quite as soon as the pragmatic responders, but our sense is that there no crucial distinction between the two groups beyond this chance occurrence.

Further research is needed to assess the validity of this account, and we cannot know whether this conclusion applies to other studies, as it would require a detailed analysis of responder-group data in each case. However, if the difference between pragmatic responders and logical responders is indeed a consequence of participants' interpretations of what is expected in a particular experimental setting, that would mean that the nature of experimental tasks/situations is of particular importance in evaluating scalar implicature processing, and must be looked at closely.

6 Conclusions

In this paper, we have discussed the results of an experiment that used visual-world eyetracking and response time measures to evaluate real-time processing of scalar implicature. In particular, we took a closer look at a result common to several studies, including ours, where participants, when faced with an underinformative utterance, tended consistently to give either a pragmatic response (reflecting scalar implicature processing) or a logical response (which did not reflect implicature processing). Other experiments have provided robust evidence suggesting that children do not process scalar implicature in underinformative situations, and that adults usually do, but not always. For the minority of adults that respond logically, it has usually been taken to mean that, like children, they have not processed the scalar implicature.

Our results suggest that the correlation between participants' responses and the generation of scalar implicatures is less straightforward. In fact, our data suggest that participants are aware of scalar implicature regardless of how they respond to underinformative sentences. We hypothesize that adults may in fact be responding according to a strategy related to the nature of the experiment, or what they expect the experimenter wants, or what a "logical" person would say — but not necessarily according to whether they themselves have actually generated the implicature.

If participants' off-line responses do not provide a reliable indication of whether or not they have generated an implicature, it appears that other tools are needed. The key would seem to be

the evaluation of off-line responses in concert with fine-grained on-line processing measures. Eye-tracking is a promising methodology for this, and we hope that future eyetracking research will further our understanding of scalar implicature processing as well as the nature of and reasons for the logical/pragmatic split between participants.

References

Bott, Lewis, and Ira Noveck. 2004. Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language* 51:433–456.

Breheny, Richard, Napoleon Katsos, and John Williams. 2006. Are generalized scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition* 100:1–30.

Foppolo, Francesca. 2007. Between "cost" and "default": a new approach to scalar implicature. In *Decalog 2007: Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue,* ed. R. Artstein and L. Vieu, 125–131. Trento, Italy.

Grice, H. Paul. 1975. Logic and conversation. In *Syntax and semantics 3: Speech acts*, ed. P. Cole & J. Morgan. New York: Academic Press.

Grodner, Daniel, Natalie Klein, Katie Carbary, and Michael Tanenhaus. 2008. Experimental evidence for rapid interpretation of pragmatic "some". CUNY conference on human sentence processing, Chapel Hill, NC.

Horn, Laurence. 1972. On the semantic properties of logical operators in English. Doctoral dissertation, UCLA.

Huang, Yi Ting, and Jesse Snedeker. 2009. On-line interpretation of scalar quantifiers: Insight into the semantic-pragmatics interface. *Cognitive Psychology* 50:376–415.

Noveck, Ira. 2001. When children are more logical than adults: Experimental investigations of scalar implicatures. *Cognition* 78:165–188.

Noveck, Ira & Andres Posada. 2003. Characterizing the time course of an implicature. *Brain and Language* 85:203–210.

Sperber, Daniel, and Deirdre Wilson. 1995. *Relevance: Communication and cognition*. Oxford: Blackwell Press, 2nd edition.

Storto, Gianluca, and Michael Tanenhaus. 2005. Are scalar implicatures computed online? In *Proceedings of Sinn und Bedeutung*, ed. E. Maier, C. Bary, and J. Huitink, 9:431–445.

Tavano, Erin, and Elsi Kaiser. To appear. The cost of being cooperative: Evidence of effort in the processing of scalar implicature. In *Proceedings of the 45th Meeting of the Chicago Linguistic Society*.

University of Southern California
Department of Linguistics
Grace Ford Salvatori 301
Los Angeles, CA 90089–1693
tavano@usc.edu
elsi.kaiser@usc.edu