Reluctant Acceptance of the Literal Truth: Eye Tracking in the Covered Box Paradigm[1]

Florian Schwarz — University of Pennsylvania
Cory Bill — CCD/Macquarie University
Jacopo Romoli — Ulster University

Abstract. Since Bott and Noveck (2004), there has been an ongoing discussion about whether scalar implicatures are delayed in online processing relative to literal meaning. Bott and Noveck (2004) provided Reaction Time evidence for such a delay, replicated in a number of later variations of their study (e.g., Bott et al., 2012). Breheny et al. (2006) found corresponding delays in self-paced reading. More recently, the issue has been investigated using the visual world paradigm, where results have been more mixed. Huang and Snedeker (2009, and subsequent work) have found delays for eye movements based on the 'not all' implicature of 'some.' But various others, (e.g., Grodner et al. (2010), Breheny et al. (2013), and Degen and Tanenhaus (2011)), report results which they argue show that implicatures are available immediately. Schwarz et al. (2015) added another angle to this picture, by using a sentence picture matching task using a Covered Picture (or Covered Box; henceforth CB; Huang et al., 2013), that allowed RT comparisons both within acceptance (target) and rejection (CB) responses. While replicating the delay for implicature-based rejection responses, they find the reverse pattern for acceptance responses, with faster RTs for implicature-compatible conditions. They propose that delays associated with literal acceptances and implicature-based rejections result from a conflict between the two possible interpretations, rather than reflecting a cost of implicature-calculation. The present experiment extends this approach beyond RTs by combining Visual World eye-tracking with the CB paradigm. The results a) are consistent with the notion that both literal and implicature interpretations are available in parallel; b) show that literal acceptances are nonetheless only provided reluctantly, presumably due to a preference for implicature meanings, in line with Schwarz et al's proposal; and c) suggest that for both literal acceptances and implicature-based rejections, there is a competition effect between the two interpretations. In addition, the RT data display an implicature-based block-priming effect, suggesting that the resolution of this conflict can be sped up through repeated task-exposure.

Keywords: scalar implicatures, processing, visual world eye tracking

1. Introduction

A major tenet in the modern study of linguistic semantics and pragmatics is that while speakers may perceive the meaning of a given utterance in its totality, it is actually a complex construct

[1]Our work on this project has benefited greatly from helpful comments and discussion by numerous people, including Dan Grodner, Yi Ting Huang, Chris Kennedy, Jesse Snedeker, Ming Xiang, and Jérémy Zehr, as well as audiences at colloquium talks at Harvard and Northwestern University and presentations at the first author's lab, XPRAG 2015 and SuB 20, which we gratefully acknowledge. Thanks to Kendra Carson for assistance in stimuli preperation and data collection. This work was supported by NSF grant BCS-1349009 to Florian Schwarz.

out of various distinct ingredients that constitute different aspects of meaning, such as literal/truth-conditional content, conversational implicatures, and presuppositions. One key theoretical question is to understand what aspects of meaning should be differentiated, and what theoretical properties they have. From the perspective of online language processing, an additional question is how these various aspects of meaning arise in real time and how they are combined into the perceived whole of utterance meaning. Pursuing both theoretical and processing perspectives in parallel, raises the additional question as to how answers to each of these questions constrain answers to the other. In the evolving enterprise of studying different aspects of meaning with methods from experimental psycho-linguistics, scalar implicatures have been the main focus of attention for over a decade. Despite the now extensive body of work on the processing of scalar implicatures, many issues remain open, with various theoretical proposals on the market and, partly, conflicting-seeming empirical results under discussion.

The present paper contributes to these ongoing debates in several novel ways. First, while the primary focus in previous work has been on the processes leading to implicature-based responses, we extend the approach by Schwarz et al. (2015) and look at the processes leading to 'literal' responses. Secondly, we introduce a novel methodology, which combines use of the Covered Box paradigm (Huang et al., 2013) with visual world eye tracking. The results a) are consistent with the notion that both literal and implicature interpretations are available in parallel; b) show that literal acceptances are nonetheless only provided reluctantly, presumably due to a preference for implicature meanings, in line with Schwarz et al's proposal; and c) suggest that for both literal acceptances and implicature-based rejections, there is a competition effect between the two interpretations. In addition, the RT data display an implicature-based block-priming effect, suggesting that the resolution of this conflict can be sped up through repeated task-exposure.

The paper is organized as follows: In the remainder of this section , we review the basic theoretical and experimental background. Next we briefly review the approach of Schwarz et al. (2015), which we take as a starting point for the present study. Section 3 presents the new experimental design and discusses its results. Section 4 concludes.

1.1. Theoretical Background

Following the seminal work by Grice (1975), conversational implicatures are commonly seen as an ingredient of the overall conveyed meaning that goes beyond what is conventionally encoded as the literal meaning of the lexical expressions involved. To illustrate the case of scalar implicatures, which we will be concerned with, the sentence in (1a) commonly conveys (1b)

(1) a. Some of the giraffes have scarves
    b. Not all of the giraffes have scarves

Crucially, this inference is not obligatorily part of what is conveyed, as illustrated by (2) - implicatures can be cancelled (or suspended):

(2) Some of the giraffes have scarves - in fact, all of them do.

This, and other hallmark properties of implicatures, is standardly captured by assuming that there are (at least) two ingredients factoring into the overall meaning of (1a), the conventionally encoded literal meaning (based on the lexical entry for *some*), and an additional inference, derived as a scalar implicature. The conjunction of the two corresponds to the commonly perceived overall meaning of (1a) that some, but not all, of the giraffes have scarves.

(3) a. **Literal meaning:** Some, and possibly all, of the giraffes have scarves.
b. **Implicature Inference:** Not all of the giraffes have scarves.

Following the seminal work by Grice (1975) and Horn (1972), the scalar implicature is derived via reasoning about certain alternative sentences that the speaker could have uttered instead. In particular, *some* is associated with a scale of alternatives that includes *all*. In a nutshell, the hearer reasons that the speaker who uttered (2) would have used the corresponding stronger *all*-statement in (4) instead if she had had good evidence for it to be true. The hearer therefore concludes that she must not think it's true (and furthermore, that it is in fact not true if he assumes she is well informed).

(4) All giraffes have scarves.

While there are many important ongoing theoretical debates about how best to implement the derivation of scalar implicatures based on such alternatives, this brief summary shall suffice for our purposes.

1.2. Experimental Background

While the original theoretical accounts of scalar implicatures were developed within philosophy of language and did not intend to make any claims about real-time processes in the human mind during language comprehension, the nascent field of experimental pragmatics has embarked on linking theoretical perspectives with accounts of implicature computation in online language processing. The derivational nature of the Gricean story lends itself to a step-wise processing account, often referred to as a 'Literal First' model (Huang and Snedeker, 2009). On this view, each time a scalar term such as *some* is encountered, the literal 'some, and possibly all' meaning is accessed

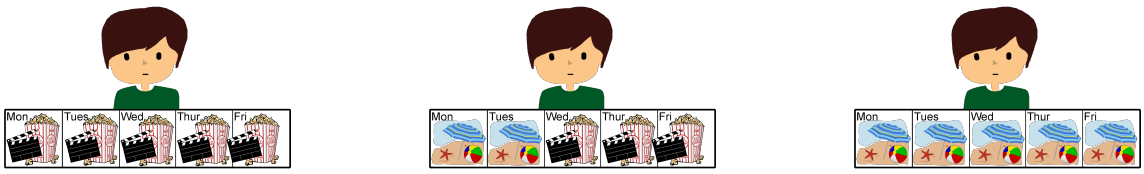
first, and the implicature inference is derived in a second step through effortful reasoning that takes time in processing.

In contrast, so-called 'Default' accounts (e.g. Levinson, 2000) are based on the notion that implicatures arise by default, and do not require effortful and time-consuming derivation. When literal interpretations are accessed at all, this happens via cancellation, which is assumed to take place in a subsequent step and thus could plausibly be assumed to take additional time in online processing.

While we're glossing over many details for reasons of space, it is clear that these two options make opposite predictions for the time-course of implicature and literal interpretations in online processing. Literal First accounts predict literal interpretations to precede implicature interpretations, whereas default accounts predict the opposite. A substantial amount of work over the last decade or so has been devoted to determining the time-course of implicature computation as compared to the computation of literal meaning. A large number of studies, starting with Bott and Noveck (2004), found implicature-based responses to be associated with a delay. Their paradigm used a truth-value judgment paradigm and presented subjects with sentences such as *Some elephants are mammals*, which could be judged either true, on a literal interpretation, or false, on an implicature interpretation, and assessed the time-course of the relevant responses in various ways. They find that implicature-based responses take more time than literal responses when subjects respond without time-constraints, and that literal responses become more frequent when the response-time window is narrowly constrained. This is argued to support a Literal First model. The general result has been replicated in a number of later variations of their study (e.g., Bott et al., 2012). Breheny et al. (2006) found corresponding delays in self-paced reading. More recently, the issue has been investigated using the visual world paradigm (Tanenhaus et al., 1995), where results have been more mixed. Huang and Snedeker (2009, and subsequent work) have found delays for eye movements based on the 'not all' implicature of 'some.' But various others, (e.g., Grodner et al. (2010), Breheny et al. (2013), and Degen and Tanenhaus (2011)), report results which they argue show that implicatures are available immediately. Given these latest results, the literature remains divided as to whether or not there is convincing support for delays being associated with implicature computation. While many other aspects of implicatures have been explored experimentally in recent years as well, such as their acquisition (Noveck 2001; Papafragou and Musolino 2003; Gualmini et al. 2001; Chierchia et al. 2001 a.o.), or their relation to other non-literal aspects of meaning (Chemla and Bott 2014; Tieu et al. 2015; Romoli and Schwarz 2015; Bill et al. 2014 a.o.), one issue that has hardly received any attention is what the processes and related time-course effects for literal interpretations of scalar sentences are. In the following section, we report an initial attempt of ours to fill this gap with a novel paradigm for assessing response times. Next, we turn to the present experiment, which extends this approach to visual world eye tracking.

2. Reaction Time Evidence for Competition between Readings

A crucial issue that arises for much of the previous work looking at response times for implicature vs. literal interpretations is that it commonly requires comparing response times for different



(Literal) (Implicature) (False)

Sentence: *Henry sometimes went to the movies this week.*

Figure 1: Target picture variants for *sometimes* conditions from Schwarz et al. (2015).

types of responses, typically True vs. False. This raises the question of whether there might be independent timing factors for the types of response, which could be separate from the nature of the underlying interpretation (literal vs. implicature) leading to it. Bott and Noveck (2004) were already aware of this and tried to guard against it in several ways. However, another approach of getting at the issue can be pursued by comparing alike responses, e.g., by assessing how easy (or fast) it is to accept statements that are only true on a literal interpretation, as compared to acceptance of statements that are also compatible with an implicature interpretation. Romoli and Schwarz (2015) utilized this approach in a picture matching task looking at implicatures (and presuppositions) under negation, and found that acceptance of pictures that were only consistent with a literal interpretation was slower than that of pictures consistent with an implicature interpretation. Schwarz et al. (2015) further generalized this approach by pursuing comparisons between both rejection and acceptance judgments and looking at implicatures in affirmative and negative sentences. As the new experiment reported below directly builds on these results, we review the latter study in some detail here.

The study used images with calendar strips representing activities of individuals on a given day of the week. A sentence and two pictures were displayed for each trial, though one was 'hidden' from view, represented by a black box. Instructions specified that only one picture would be a match for the displayed sentence, so that subjects should choose the hidden picture precisely if the visible picture did not match the sentence. Figure 1 illustrates the visible picture variants and an example sentence. Crucial comparisons concerned response times for acceptance of the Literal picture and the Implicature picture on the one hand, and rejections of the Literal and the False picture on the other. For the latter, the results displayed the standard pattern, where implicature-based rejections take longer than ones where rejection is possible based on literal meaning alone. This is in line with Bott & Noveck's findings, although it makes a different comparison by focusing on rejection responses for both cases. For acceptances, the pattern was in line with that found by Romoli and Schwarz (2015) for implicatures under negation (which also yielded the same pattern of results in the Schwarz et al. 2015 version). These effects were reflected in statistical interactions between response type and picture condition, as well as simple effects between pictures within acceptances and rejections respectively.

Figure 2: Response times for Covered Box (CB) and Target choices by condition from Schwarz et al. (2015).

In sum, rejections that can be based on literal meaning alone are faster than rejections that require the implicature, but acceptances of pictures consistent with the implicature are faster than acceptances of pictures that are only compatible with the literal interpretation. Schwarz et al. (2015) note that this cannot be accounted for solely in terms of delays for implicature computation. The delay in accepting literal pictures is unexpected from that perspective. A straightforward descriptive generalization about these results, however, is that delays arise precisely in those cases where more than one response is possible, in this case, the Literal picture. That is, accepting the Literal picture is slower than accepting the Implicature picture, and rejecting it is slower than rejecting the False picture. Given this pattern, a plausible hypothesis is that the delays in response times are due to a competition effect of sorts. Schwarz et al. (2015) frame this in terms of opposing pressures: on the one hand, implicature interpretations are generally strongly preferred, and in fact are often perceived to be THE meaning of *some*-statements by naive speakers (as can be confirmed by anyone teaching the existential quantifier $\exists$ to students of Predicate Logic). On the other hand, it is commonly assumed that some form of Charity principle plays a strong role in language comprehension, which leads hearers to try to adopt an interpretation that makes the speaker's utterance true (in our paradigm, this means to prefer the visible picture to match the sentence). These two pressures oppose one another in the Literal condition, in that the first favors rejecting the visible picture (i.e., selecting the covered box), whereas the second favors accepting the visible picture. A plausible interpretation of the response time delays then is to see them as due to having to resolve the conflict between these two options, i.e., as a competition effect of sorts. Such an account, which extends to other response time results in the literature, does not need to allude to delays in implicature computation, though it is also not incompatible with such delays (as we will discuss further when looking at the current experiment's results). To the extent that these are supported by more fine-grained online processing results, such as from visual world eye tracking, it might well be the case that implicatures arise after literal meaning, leading to the type of effect found by Huang and Snedeker (2009), but that response time effects, which commonly are larger than those found in eye tracking, are at least in part, or perhaps even entirely, due to the competition effect described here. The competition effect itself can also be further elucidated by using the more

fine-grained information that eye tracking affords. This is exactly what the experiment reported below aims to achieve.

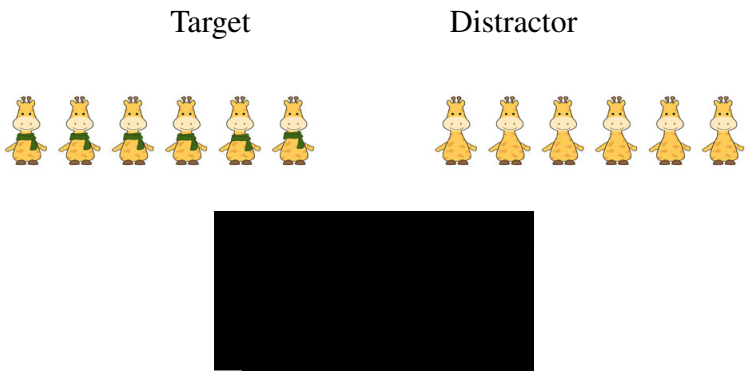
3. A Novel Paradigm: Visual World Eye Tracking with a Covered Box

The visual world eye tracking paradigm (Tanenhaus et al., 1995) provides eye movement data relative to a visual scene as a sentence unfolds. It is commonly paired with a picture selection task, where subjects have to identify a picture as a match for the description provided by the sentence. As it provides continuous data as the sentence unfolds, it allows more direct insights into the processes involved in arriving at the final picture selection. This has been used in the study of scalar implicature to great effect (starting with Huang and Snedeker, 2009). Previous work in this area followed the standard approach of focusing on a temporary ambiguity: a sentence such as *Point to the girl that has some of the soccer balls.*, is presented along with pictures where one girl has some, but not all of the soccer balls and another has all of the socks. Up until the second part of *soccer*, a literal interpretation is consistent with either individual, whereas an implicature interpretation is only consistent with the girl with the soccer balls.

The covered box paradigm of Schwarz et al. (2015) allows for the presentation of items where: a) a real choice remains to be made at the end of the sentence, and b) the visible picture candidate for matching the relevant sentence is only consistent with its literal interpretation. Therefore, by combining this approach with the visual world paradigm, we gain further insights into the processes involved in reaching both acceptance and rejection interpretations for both implicature and literal interpretations, which in turn allows us to test more directly for the competition effects posited by Schwarz et al. (2015). To capture any potential effect of having to make a real choice at the end of the sentence, and in particular, of considering visible pictures only consistent with a literal interpretation, we added conditions where the ambiguity was only temporary, as in previous visual world tasks. That is, in these trials the ultimate choice was always the covered box as the visible pictures were incompatible with the literal meaning of the sentence once it unfolded in its entirety. By using a block design with counterbalanced block-order, we were able to investigate effects of globally ambiguous trials on online processing and behavioral response measures, specifically response times.

3.1. Materials & Design

Each trial involved 3 pictures, a target, a distractor, and the covered box, along with an auditorily presented sentence, as illustrated in Figure 3. The sentence always had the format *Some of the* NP_1 *have* NP_2, where 'NP_1' was an animal name and 'NP_2' an accessory. The distractor was never a live option for being a match at any point during the sentence, and was only included to provide a richer visual scene and promote visual exploration (distractor-like pictures were the correct choice on certain filler trials; see below). The covered box was introduced as hiding yet another picture



Sentence: *Some of the giraffes have scarves.*

Figure 3: Illustration of trial display in Literal-Global condition.

Implicature-Global Literal-Global Implicature-Temp Literal-Temp

Figure 4: Illustration of target picture variations by condition.

in the instructions, which was to be chosen if and only if none of the visible pictures matched the picture. The target picture was varied along two dimensions, as shown in Fig. 4. First, it either was consistent with the implicature or not, which we encoded as IMPLICATURE vs. LITERAL: in the former variant, only some of the giraffes had a given accessory, whereas in the latter, all did. Secondly, the accessory mentioned in the sentence (e.g., scarf, in Fig. 3) either did or did not match the one displayed in the target picture. When it did (e.g., the giraffes have scarves), the target remained a candidate referent after the sentence was complete (modulo the implicature vs. literal interpretation subjects adopted). When it did not (e.g., when the giraffes wore hats), the covered box was the only viable final response choice, though prior to the sentence-final noun, the target remained a candidate referent (again, modulo the implicature vs. literal interpretation issue). This factor was encoded as GLOBAL vs. TEMP(ORARY), and was blocked in the experimental materials to allow assessment of the influence the two types of trials might have on one another (see details under Procedures below).

A total of 16 experimental items was created, each with variants in the 4 conditions, which were then divided into 4 counter-balanced lists, each with 4 items per condition. Several types of fillers were included as well to conceal the experimental manipulation and to counterbalance various potential issues that could arise from the experimental stimuli:

- 8 fillers with sentences with *some*, parallel to the experimental stimuli, 4 of them with a target picture that was consistent with the implicature, and 4 with a target-like picture (i.e., only some of the animals had a given accessory) but with an accessory different than that mentioned in the sentence, leading to a covered box choice.
- 16 fillers using *none* as the determiner, with pictures that were structurally identical to a counter-balanced set of target pictures, but with the picture corresponding to the distractor as the correct answer choice.
- 16 fillers with definite plurals for NP_1 (e.g., *The giraffes...*), with a target picture where all animals had a given accessory, which matched in 8 fillers and did not match in the other 8.

This selection of fillers ensured a relatively balanced distribution of visible picture vs. covered box choices overall, and within the former, of target-type vs. distractor-type picture choices. Note also that there were no items using the universal quantifier *all*, as we used plural definites instead. Finally, note that there were only 4 items where it ultimately mattered whether or not subjects adopted an implicature or literal interpretation for their response.

3.2. Participants & Procedure

Seventy-eight undergraduate students at the University of Pennsylvania took part in the experiment for course credit. They were seated in front of a monitor and had their eye movements recorded using an EyeLink 1000 eye tracker by SR Research. Each session began with instructions describing the task as involving the selection of a picture, via mouse click, as a match for a description provided by a sentence. Furthermore, the nature of the covered box was explained, and two practice trials with plural definites illustrated when the a visible picture would be chosen and when the covered box would be chosen. Subjects could then ask questions about the task, if any, and the eye tracker was set up. Each subject was assigned to a group that would see one of the counterbalanced lists, which included a block order manipulation, such that half the subjects saw the global conditions first, and the other half the temporary conditions first.

3.3. Results

3.3.1. Responses and Response Patterns

Response accuracy in the unambiguous conditions (both TEMPORARY and IMPLICATURE-GLOBAL) was at ceiling ($> 98\%$ Covered Box and Target choices respectively). The LITERAL-GLOBAL condition yielded 22.5% literal target responses, and 77.5% implicature-based CB responses.

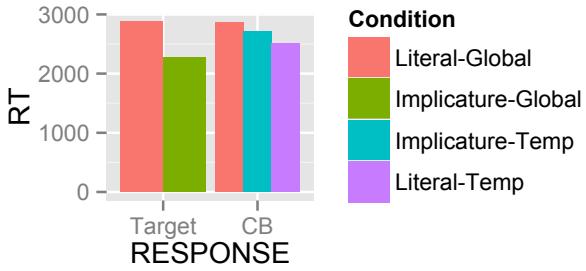


Figure 5: Reaction times by condition and Response.

Turning to the response time data, only trials with appropriate responses (i.e., accurate responses in unambiguous conditions and either response in the LITERAL-GLOBAL condition) were included in the statistical analyses, which used mixed effect models with maximal random effect structures (Barr et al. 2013). The crucial conditions are illustrated by response in Figure 5. A first comparison, parallel to that in Romoli and Schwarz (2015); Schwarz et al. (2015), between target responses in the LITERAL-GLOBAL and IMPLICATURE-GLOBAL Target responses, replicated the previous findings, in that the former are significantly slower than the latter (2887ms vs. 2275ms; $\beta = -599.4$, $SE = 145.4$, $t = -4.12$, $\chi^2 = 13.29$, $p < .001$).

Secondly, covered box choices in the LITERAL-GLOBAL condition were slower than in the LITERAL-TEMP condition (2858ms vs. 2511ms; $\beta = -408.5$, $SE = 164.8$, $t = -2.48$, $\chi^2 = 5.35$, $p < .05$), which is in line with delays for implicature-based rejections standardly reported in the literature: in the latter cases, responses are based on the literal meaning of the final noun phrase (*scarves*), which does not match the accessories in the picture, while in the former case, the only grounds for rejection is an implicature interpretation.

Interestingly, covered box choices in the IMPLICATURE-TEMP condition were also slower than in the LITERAL-TEMP condition (2707ms vs. 2511ms; $\beta = -200.36$, $SE = 80.38$, $t = -2.49$, $\chi^2 = 5.10$, $p < .05$), even though in both cases, the literal meaning suffices to reject the target picture, due to the mismatch between accessories in the sentence and picture. This seems to indicate that temporary compatibility with the implicature affects RTs even when the sentence as a whole is incompatible with the picture based on literal meaning alone.

Finally, considering the impact of the block-order factor (i.e., whether GLOBAL or TEMP trials were seen first), there was a significant effect of order, in that covered box choices in the LITERAL-TEMP condition were faster when they had been preceded by the block of GLOBAL trials. This suggests that an implicature interpretation aided to speed up the rejection of the visible picture in the LITERAL-TEMP condition, but only after previous exposure to GLOBAL trials.

To summarize the response time findings, we replicate the previous results from Romoli and

Schwarz (2015) and Schwarz et al. (2015) in that target choices are faster when the target is compatible with an implicature interpretation. Secondly, we find that rejection of a visible picture purely on the grounds of an implicature interpretation is slower than rejection based on literal meaning, again in line with previous results from Schwarz et al. (2015), though with a slightly different literal-based rejection comparison provided by the TEMP condition. The finding that the IMPLICATURE vs. LITERAL manipulation even had an effect in the TEMP condition is a novel and somewhat surprising finding. The temporary viability of the picture in light of an implicature interpretation seems to make the picture attractive enough to slow down rejection overall, even though rejection can be based on literal meaning alone once the final noun is heard. The block order effect speaks to a similar point, in that previous exposure to GLOBAL trials speeds up rejection in the LITERAL-TEMP condition, which again suggests that an implicature interpretation is at play even in that condition, and that this is facilitated, and perhaps sped up, by previously having seen GLOBAL trials, where the existence of both interpretations is highlighted and a choice between them is forced. We return to this point in the general discussion below.

3.3.2. Eye tracking

For the purposes of analyzing the eye movement data, we computed Target Advantage scores, which here were defined as looks to target – looks to other pictures, time-locked to the auditory onset of crucial expressions, specifically the sentence-initial quantifier (see Fig. 6) and the sentence-final noun (see Fig. 7). Statistical analyses using mixed effect models were conducted on Target Advantage scores transformed to Elogit for 200ms time windows after the relevant onsets.

Starting with the distribution of looks in the GLOBAL conditions (independent of response, effectively pooling the two red lines in the graph in Fig. 6) there is a relative decrease in looks to the target in the LITERAL condition, marginally significant in the 800-1000ms time window and fully significant in the 1000-1200ms time window ($\beta = 1.90$, $SE = 050$, $t = 3.82$, $\chi^2 = 14.47$, $p < .001$). This is comparable to the time-course of implicature-based looking patterns in Huang and Snedeker (2009) and following work.

Given our interest in the emergence of responses based on a literal interpretation, we next turn to a comparison in fixation patterns in IMPLICATURE and LITERAL trials in the GLOBAL conditions that ended in selection of the target (see Fig. 6a). From 1400-1800ms after Q-Onset, there is a significant relative decrease in looks to the target in the LITERAL condition (1400-1600ms: $\beta = 3.03$, $SE = 1.16$, $t = 2.61$, $\chi^2 = 5.12$, $p < .05$; 1600-1800ms: $\beta = 3.56$, $SE = 1.15$, $t = 3.08$, $\chi^2 = 7.08$, $p < .01$), and this effect is still nearly marginally significant from 1800-2000ms ($\beta = 2.57$, $SE = 1.51$, $t = 1.70$, $\chi^2 = 2.57$, $p = .11$). Comparing GLOBAL-LITERAL trials based on the ultimate response (Fig. 6b), a difference in fixation patterns only emerges around 2000ms (1800-2000ms: $\beta = 1.27$, $SE = 0.73$, $t = 1.74$, $\chi^2 = 2.71$, $p < .1$; 2000-2200: $\beta = 2.25$, $SE = 0.61$, $t = 3.69$, $\chi^2 = 9.43$, $p < .01$). In sum, LITERAL target acceptance trials display a

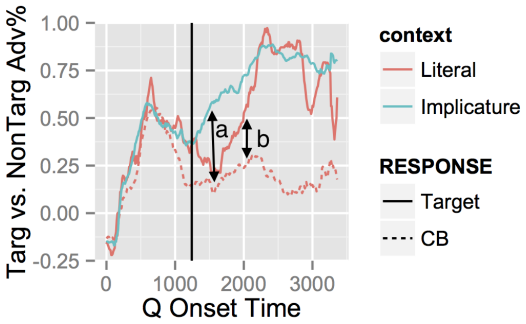


Figure 6: Target Advantage plot for **Global** trials, time-locked to the onset of the quantifier. The vertical black bar indicates the average onset of the final noun.

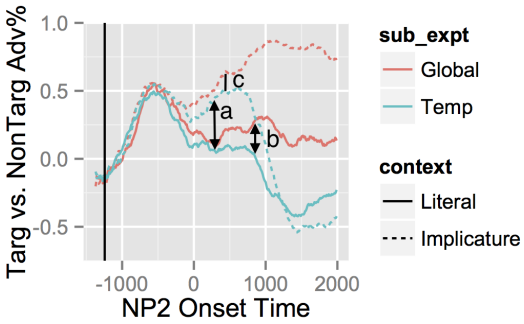


Figure 7: Target Advantage plot for Target (**Global-∃-Pic**) and CB choice trials (all others) by condition, time-locked to the final noun phrase. The vertical black bar indicates average onset of the quantifier.

phase of disfavoring the Target, compared to IMPLICATURE acceptance trials, and pattern together with LITERAL target rejection trials up until 2000ms after quantifier onset.

Turning to a comparison of the time course in GLOBAL and TEMP conditions with both LITERAL and IMPLICATURE picture versions, we find further evidence for implicature interpretations being at play even when this turns out to be innocuous for the final response. Focusing on trials where the covered box is chosen for the TEMP conditions as well as the LITERAL-GLOBAL condition (and on target selection trials for the IMPLICATURE-GLOBAL condition), Fig. 7 displays Target Advantage scores relative to the onset of the sentence-final noun. Interestingly, the IMPLICATURE-TEMP condition displays a sustained relative increase in looks to the target relative to the LITERAL-TEMP condition even after the onset of the noun (which, recall, does not match the accessory in the picture in the TEMP conditions (Fig. 7a). This is significant in 200ms time windows from 200-1000ms after the final noun (e.g., 200-400ms: $\beta = 4.39$, $SE = 0.99$, $t = 4.41$, $\chi^2 = 10.62$, $p < 0.1$). This suggests that the target is discarded relatively quickly based on the implicature interpretation in the LITERAL-TEMP condition, whereas it continues to be scrutinized for its relation to the sentence-final noun in the IMPLICATURE-TEMP

Furthermore, even in the LITERAL-GLOBAL trials that end in a covered box-choice, participants continue to look at the target much more than in LITERAL-TEMP trials from around 800ms after the onset of the final noun (Fig. 7b; 600-800ms: $\beta = 1.89$, $SE = 1.04$, $t = 1.82$, $\chi^2 = 2.78$, $p < 0.1$; 800-1000ms: $\beta = 3.93$, $SE = 0.97$, $t = 4.05$, $\chi^2 = 8.82$, $p < 0.01$). This suggests that even when the covered box is ultimately chosen based on an implicature interpretation, there is a lingering awareness of the potential viability of the LITERAL target picture version. The IMPLICATURE-GLOBAL and IMPLICATURE-TEMP fixation patterns come apart even earlier (Fig. 7c), at 400ms after the final noun ($\beta = 1.53$, $SE = 0.71$, $t = 2.16$, $\chi^2 = 4.15$, $p < 0.05$),

suggesting that the literal incompatibility of of the final noun phrase in the TEMP condition has an impact relatively quickly.

In sum, the path to target acceptance responses in the LITERAL-GLOBAL condition as reflected in fixation patterns differs from that in the IMPLICATURE-GLOBAL condition, in that there is a temporary decrease in looks to target in the latter. Furthermore, both IMPLICATURE-TEMP trials and LITERAL-GLOBAL trials display increased looks to the target, suggesting the active consideration of an implicature interpretation in the former and of a literal interpretation in the latter.

3.4. Discussion

This experiment combined two established psycholinguistic techniques (covered box and visual world) to create a novel paradigm for exploring the processing of scalar implicatures. It was designed to be a natural followup to Schwarz et al. (2015), and continue investigating the suggestion that the response time delays, traditionally attributed to scalar implicature computation, might instead be a result of participants' grappling with with the 'opposing pressures' pushing them towards different interpretations (literal and implicature) of the relevant sentences. If this opposing pressure hypothesis is on the right track, we would expect to see it revealed through a replication of Schwarz et al. (2015)'s response time patterns, as well as through eye-movement patterns that indicate participants are considering both interpretations in parallel at some point prior to responding.

3.4.1. Response Times

First, let us consider the predictions these theories make with regard to response times. The proposal that literal interpretations are considered first and implicature computation comes with a delay predicts increased response times for responses that require an implicature interpretation, relative to ones that are based only on a literal interpretation. The opposing pressures hypothesis, on the other hand, expects response-time delays to appear whenever participants are presented with a condition in which both the literal and implicature interpretations are in competition.

Looking to our response time results, we found that participants took longer to reject target pictures that were only consistent with the literal interpretation (LITERAL-GLOBAL), than to reject target pictures that were only temporarily consistent with the literal interpretation (LITERAL-TEMP). This pattern is consistent with the idea that implicature computation is delayed, as participants are taking longer to provide responses that require computation of an implicature (namely rejections in the LITERAL-GLOBAL condition), than to provide responses that could be based only on the literal meaning (rejecting LITERAL-TEMP). However, this pattern is also consistent with the opposing

pressures hypothesis, given that both literal and implicature interpretations remain live options in the LITERAL-GLOBAL condition right up until a choice is made between them by the participant, whereas in the LITERAL-TEMP condition the conflict is resolved once the final noun makes the target picture false according to both interpretations. Therefore, it seems that this difference in response times does not favor either explanation over the other.

We also found that participants took longer to select target pictures only consistent with the literal interpretation (LITERAL-GLOBAL), than to select target pictures also consistent with the implicature interpretation (IMPLICATURE-GLOBAL). This pattern is a replication of equivalent conditions from Romoli and Schwarz (2015) and Schwarz et al. (2015), and is more informative towards the aim of distinguishing between the two theories of interest, as the implicature delay hypothesis does not predict such a pattern. That is, we would expect at least some of the responses in the IMPLICATURE-GLOBAL condition to involve an implicature interpretation, while the target picture selections in the LITERAL-GLOBAL condition can only be based on a literal interpretation. Thus, according to the implicature delay hypothesis, the latter thus should be faster, if anything, but we find the opposite. In contrast, the opposing pressures hypothesis can account for this pattern straightforwardly, by noting that the LITERAL-GLOBAL condition forces participants to make a choice between the two possible interpretations, whereas, the IMPLICATURE-GLOBAL condition does not present such a conflict, with both the literal and implicature interpretations leading to selection of the target picture. Therefore, it seems that this pattern of response times is only consistent with the predictions of the opposing pressures hypothesis. (But see below for discussion of the possibility of a possible way to reconcile these results.)

Another important result was that we found rejections of pictures that were temporarily consistent with an implicature interpretation (IMPLICATURE-TEMP) to take longer than rejections of pictures that were temporarily only consistent with a literal interpretation (LITERAL-TEMP). This is a novel result, and seems to suggest that participants are more reluctant to abandon a target picture consistent with the implicature interpretation, than a target picture consistent only with the literal interpretation. This pattern is not predicted by either of the theories of interest, and so doesn't seem to be of use in their evaluation. However, it is notable that this pattern is consistent with an idea considered by Schwarz et al. (2015) (see also discussion by various previous authors, perhaps most prominently Levinson, 2000), that the implicature interpretation is the preferred/default interpretation of these sentences. After all, consideration of the implicature turns out to be entirely useless by the end of the sentences in this condition, and yet we find a strong impact of the implicature interpretation on behavioral response variables.

Finally, there was a block order effect in the present results, such that participants who had been presented with the GLOBAL trials first were faster in rejecting the target pictures in the LITERAL-TEMP condition. This suggests that consideration of trials where deciding on a response rides on deciding between an implicature and a literal interpretation has an effect on the processing of implicature interpretations in relation to literal ones later on in the experiment. More specifically, having had to decide between implicature and literal interpretations earlier on speeds up the de-

cision to respond based on an implicature later on in the TEMP conditions, prior to encountering the final noun which makes the literal meaning of the sentence incompatible with the picture in these conditions. This effect could be accounted for in different ways by the two approaches under consideration. From the perspective of the opposing pressures hypothesis, a way of capturing this priming effect would be to say that people become more practiced in making a decision between two available interpretations, and thus can make their judgment based on that decision more quickly in the TEMP conditions than if they had to wait until they can base their response on the final noun. The proposal that implicature computation induces a delay, on the other hand, could see this as a priming effect on the computation of an implicature interpretation, so that this interpretation becomes available more quickly after some practice. Note, however, that this practice effect must arise quite quickly, as there only is a total of 4 experimental trials where the response crucially depends on deciding between an implicature interpretation and a literal interpretation. The present response time data do not distinguish between these two options, but the effect should be of some interest for further study. Future work should also investigate whether there are comparable reflexes of such a priming effect in fixation patterns in the early phases of interpretation after encountering the quantifier.

In sum, it would appear that where the response time patterns distinguish between the two theories of interest, they seem to support the opposing pressures approach over the implicature delay theory. Notably, none of these results actually 'rule-out' the possibility that there is a delay in implicature computation, which is then further extended by the opposing pressures competition effect. However, without independent evidence showing that implicature computation is indeed contributing to the delay, considerations of parsimony lead us to prefer attributing the delay to opposing pressures explanation alone. We mention this because the eye tracking results that we turn to next will change the situation somewhat.

3.4.2. Eye tracking

First, as before, let us consider the relation of the theories under investigation to the eye movement results. On a simple two-stage version of the implicature delay theory, which assumes that there's an initial phase where only the literal interpretation is considered and a second phase where only the implicature interpretation is considered, participants' eye movements are expected to reflect these two stages: first, eye movements should be attracted to any pictures compatible with the literal interpretation, and subsequently, attention should be restricted to those compatible with the implicature interpretation. On the other hand, the opposing pressures approach would predict there to be a phase where participants' eye movements oscillate as the two interpretations are weighed against one another.

Considering trials ending in target picture selections in the IMPLICATURE-GLOBAL and LITERAL-GLOBAL conditions first, the latter exhibit a phase with a relative decrease in looks to the target,

despite being ultimately chosen. This is surprising on a two-stage account with a delay for implicatures, as literal responders would not be expected to go beyond the initial, literal phase in the first place. But if there is a phase where both interpretations are considered in parallel, as on the opposing pressures approach, this back and fourth in the LITERAL-GLOBAL condition is very much expected, regardless of what final interpretation is being adopted. And in fact, the results for trials in the LITERAL-GLOBAL condition ending in a covered box choice further support this point: relative to the LITERAL-TEMP condition, the proportion of looks to the target continues to be higher in the final trial-phase, suggesting that even when subjects respond based on the implicature interpretation, they continue to consider the literal interpretation as a potential alternative.

Another interesting aspect of the eye-tracking results was revealed when comparing the LITERAL-TEMP and IMPLICATURE-TEMP conditions. Given the mismatch of the mentioned noun and the accessory in the target picture, the ultimate choice here was the covered box. Nonetheless, we found a difference in fixation patterns, in that the IMPLICATURE-TEMP condition showed sustained looks to the target picture after the onset of the final noun. This pattern provides more insight into the relative delay participants displayed in their response times in the IMPLICATURE-TEMP condition, compared to the LITERAL-TEMP condition. In particular, it lends further support to the notion that the response time delay is due to the temporary compatibility of the picture with an implicature interpretation, as the picture gets scrutinized further for compatibility with the noun in the IMPLICATURE-TEMP condition, while it is abandoned more quickly in the LITERAL-TEMP condition, which can be rejected before hearing the noun if and only if an implicature interpretation is adopted. Thus, that interpretation seems to be adopted relatively quickly, before the sentence final noun comes in, consistent with the notion that implicature interpretations are preferred, and may even have the status of being the default choice.

What our eye movement clearly show, then, is that there is a phase where both implicature and literal interpretations are considered, and that responses based on literal interpretations generally do involve consideration of the implicature interpretation at some point. Furthermore, they suggest that implicature interpretations do constitute the default, in that they give rise to effects even when they ultimately turn out to be irrelevant, as in the TEMP conditions. Overall, this supports the interpretation of the response time results in Schwarz et al. (2015) in terms of opposing pressures favoring the respective interpretation options. They are clearly incompatible with the simple two-stage implicature delay model as outlined above. Note, however, that that is not say that they are altogether incompatible with the notion that implicature computation takes time. An alternative account could consider a three-stage variant, where an initial literal phase is followed by a phase where the implicature interpretation is considered along with the literal interpretation, which can be followed by a decision for the literal interpretation - in line with the notion of implicature cancellation posited by default accounts. In effect, this would combine the competition aspect of the opposing pressures approach with an initial implicature delay (and more generally, different aspects of Literal First and Default accounts). Indeed, one aspect of our data could well be seen as supporting a delay account, as we see a decrease in looks to the target in the LITERAL-GLOBAL condition at a time-point (1000-1200ms) that is comparable to the implicature-based looking pat-

tern found by Huang and Snedeker (2009). This decrease in looks to the target in the LITERAL-GLOBAL condition could be interpreted as being indicative of the point at which participants move from accessing the initial literal interpretation to the more costly implicature interpretation.

While this combination of an initial implicature delay and a competition phase is consistent with our data, it is in principle also possible to try to account for initial delay effects in terms of an emerging competition effect. In particular, it could be that it is not access to an implicature interpretation itself that is delayed, but rather the emergence of the preference for that interpretation. In effect, the opposing pressures account could posit a temporal asymmetry for the two pressures at play, with a charity-like preference for true interpretation being operative early on, and a preference for implicature interpretations arising later on. As far as we can tell, existing data does not clearly distinguish between these two possibilities.

4. Conclusion

In this paper, we use a novel paradigm combining Visual World eye-tracking with the Covered Box paradigm to investigate the time course of literal and implicature interpretations. In particular, we aimed at investigating the proposal by Schwarz et al. (2015) that the delays which arise with sentences involving scalar implicatures is at least in part due to a competition effect between different possible interpretations. The results suggest that there is a phase where both literal and implicature interpretations are available in parallel. Moreover they show that (at least in this paradigm) literal acceptances are only provided reluctantly, after consideration of an implicature interpretation, presumably due to a preference for implicature meanings, in line with Schwarz et al.'s proposal and more generally the notion that implicature interpretation constitute a default of sorts. They thus support the idea that there is a competition effect between the two potential interpretations, which is present for both literal acceptances and implicature-based rejections. In addition, the response time data display an implicature-based block-priming effect, suggesting that the resolution of this conflict can be sped up through repeated task-exposure.

References

Bill, C., J. Romoli, F. Schwarz, and S. Crain (2014). Scalar implicatures versus presuppositions: The view from acquisition. *TOPOI, Special Issue on "Presuppositions: Philosophy, Linguistics, and Psychology"*, 1–15.

Bott, L., T. M. Bailey, and D. Grodner (2012). Distinguishing speed from accuracy in scalar implicatures. *Journal of Memory and Language 66*(1), 123–142.

Bott, L. and I. A. Noveck (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of memory and language 51*(3), 437–457.

Breheny, R., H. J. Ferguson, and N. Katsos (2013). Investigating the timecourse of accessing conversational implicatures during incremental sentence interpretation. *Language and Cognitive Processes 28*(4), 443–467.

Breheny, R., N. Katsos, and J. Williams (2006, July). Are generalised scalar implicatures generated

by default? an on-line investigation into the role of context in generating pragmatic inferences. *Cognition 100*(3), 434–463.

Chemla, E. and L. Bott (2014). Processing inferences at the semantics/pragmatics frontier: disjunctions and free choice. *Cognition 130*(3), 380–396.

Chierchia, G., S. Crain, M. T. Guasti, and R. Thornton (2001). 'Some' and 'or': A study on the emergence of logical form. In S. C. Howell, S. A. Fish, and T. Keith-Lucas (Eds.), *Boston University Conference on Language Development (BUCLD) 24*, Somerville, Massachusetts, pp. 22–44. Cascadilla Press.

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

Grice, H. (1975). Logic and conversation. In P. Cole and J. L. Morgan (Eds.), *Syntax and Semantics*, Volume 3, pp. 41–58. New York: Academic Press.

Grodner, D. J., N. M. Klein, K. M. Carbary, and M. K. Tanenhaus (2010). "Some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition 116*(1), 42–55.

Gualmini, A., S. Crain, L. Meroni, G. Chierchia, and M. T. Guasti (2001). At the semantics/pragmatics interface in child language. In R. Hastings, B. Jackson, and Z. Zvolenszky (Eds.), *Semantics and Linguistic Theory (SALT) 11*, Cornell University, Ithaca, NY, pp. 231–247. CLC Publications.

Horn, L. (1972). *On the Semantic Properties of Logical Operators in English*. Ph. D. thesis, UCLA.

Huang, Y., E. Spelke, and J. Snedeker (2013). What exactly do number words mean? *Language Learning and Development 9*(2), 105–129.

Huang, Y. T. and J. Snedeker (2009). Online interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cognitive psychology 58*(3), 376–415.

Levinson, S. C. (2000). *Presumptive meanings: the theory of generalized conversational implicature*. Language, speech, and communication. Cambridge, Mass: MIT Press.

Noveck, I. (2001). When children are more logical than adults: experimental investigations of scalar implicatures. *Cognition 78*(8), 165–188.

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

Romoli, J. and F. Schwarz (2015). An experimental comparison between presuppositions and indirect scalar implicatures. In F. Schwarz (Ed.), *Experimental Perspectives on Presuppositions*, pp. 215–240. Springer International Publishing.

Schwarz, F., J. Romoli, and C. Bill (2015). Scalar implicature processing: Slowly accepting the truth (literally). In *Proceedings of Sinn und Bedeutung 19*, pp. 553–570.

Tanenhaus, M. K., M. J. Spivey-Knowlton, K. M. Eberhard, and J. C. Sedivy (1995). Integration of visual and linguistic information in spoken language comprehension. *Science 268*(5217), 1632–1634.

Tieu, L., J. Romoli, P. Zhou, and S. Crain (2015). Children's knowledge of free choice inferences and scalar implicatures. *Journal of Semantics (Advance Access)*.