Cognitive Psychology 58 (2009) 376–415

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface

Yi Ting Huang *, Jesse Snedeker

Department of Psychology, William James Hall, Harvard University, Cambridge, MA 02138, USA

ARTICLE INFO

Article history:
Accepted 2 September 2008
Available online 1 November 2008

Keywords:
Semantics
Pragmatics
Scalar implicature
Quantifiers

ABSTRACT

Scalar implicature has served as a test case for exploring the relations between semantic and pragmatic processes during language comprehension. Most studies have used reaction time methods and the results have been variable. In these studies, we use the visual-world paradigm to investigate implicature. We recorded participants' eye movements during commands like "Point to the girl that has some of the socks" in the presence of a display in which one girl had two of four socks and another had three of three soccer balls. These utterances contained an initial period of ambiguity in which the semantics of *some* was compatible with both characters. This ambiguity could be immediately resolved by a pragmatic implicature which would restrict *some* to a proper subset. Instead in Experiments 1 and 2, we found that participants were substantially delayed, suggesting a lag between semantic and pragmatic processing. In Experiment 3, we examined interpretations of *some* when competitors were inconsistent with the semantics (girl with socks vs. girl with no socks). We found quick resolution of the target, suggesting that previous delays were specifically linked to pragmatic analysis.

1. Introduction

Where does language end and communication begin? While many aspects of utterances are tightly linked to word meaning and syntactic structure, other facets are clearly added by context-sensitive, inferential processes. For example, in the dialogue in (1), we can infer from the response of the Little

* Corresponding author. Fax: +1 617 496 0975.
E-mail addresses: huang@wjh.harvard.edu (Y.T. Huang), snedeker@wjh.harvard.edu (J. Snedeker).

0010-0285/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.cogpsych.2008.09.001

Red Hen that she has not finished making the bread. Indeed, if we do not make this inference her response would be a non-sequitur.

(1) The Lazy Dog: Have you made the bread yet?
The Little Red Hen: I've planted the grain.

But this inference is not part of the truth conditional content of the Hen's statement. Planting the grain does not rule out the possibility of making bread. In fact on the farm, one event typically precedes the other. This division between semantically-encoded meaning and the inferences that we can derive from it was made prominent by Grice (1957, 1975). Semantics is used to refer to the truth conditional content of the utterance or the aspects of the interpretation that can be directly calculated from the meanings of words and the structural relationships between them. In contrast, pragmatics is used to refer to the aspects of interpretation arrived at by an inferential analysis of the utterances with respect to the context and the communicator's goals. Grice proposed that while pragmatic inferences make use of the semantic analysis, they are distinct from this truth conditional content because they are in fact defeasible. In other words, we can imagine a situation where our initial inference (the bread is not finished) is explicitly canceled by subsequent statements from the Hen in (2).

(2) The Little Red Hen: (*frustrated*) And I've cut the wheat and ground the flour. In fact, I've done everything and now I will eat the bread all by myself!

While Grice's distinction between semantic content and pragmatic inference has been widely accepted, there are divergent theories about the nature of these two levels of representation and their relation to one another (Levinson, 1983, 2000; Recanati, 2003; Sperber & Wilson, 1986/1995). These theories are not psycholinguistic models but they differ in their conception of the processes that mediate between semantic and pragmatic interpretation and how they might interact. These processes have been explored by psycholinguists, primarily by examining the effects of context on language comprehension or contrasting the processing of utterances that require a particular inference with those that do not. Evidence of early pragmatic processing has been demonstrated across phenomena as diverse as the resolution of lexical ambiguity, the use of contrast sets to predict the referent of a modified noun, and the interpretation of metaphoric expressions (Frisson & Pickering, 1999; Glucksberg, Gildea, & Bookin, 1982; Rayner & Duffy, 1986; Sedivy, Tanenhaus, Chambers, & Carlson, 1999). For example, Sedivy and her colleagues (1999) demonstrated that listeners were quicker to comprehend "Pick up the tall glass" in the presence of another contrasting member of the same category (e.g., a short glass). This rapid sensitivity to the presence of a pragmatically specified comparison set suggests that upon hearing *tall*, listeners were able to quickly generate an inference that the referent is likely to belong to a set of objects from the same category (presumably the one with a short and tall item) (Grodner & Sedivy, in press; Sedivy, 2003).

However, this research leaves open the question of whether these rapid pragmatic inferences are preceded by some degree of semantic analysis. In fact in many cases, pragmatic inferences seem to depend upon aspects of lexical and compositional semantics. In the example above, the relevance of the contrast set can only be determined after recognizing that *tall* is a scalar adjective that encodes the dimension of height. Thus we might expect to find some moment in processing—however brief it may be—when the semantic contribution a given word is available but the pragmatic inference that it triggers is not. The experiments in this paper are an attempt to find that moment. We do this by exploring a test case where the division between semantic meaning and pragmatic inference is sharply defined: the interpretation of scalar quantifiers.

1.1. The Neo-Gricean theory of scalar interpretations

Linguists have long noted that terms like *some* have two distinct interpretations (Horn, 1972, 1989; Gadzar, 1979). Typically, sentences like (3) will be taken to imply that Ernie ate only a proper subset of the apples (he did not eat all of them).

(3) Bert: Where are the apples that I bought?
Ernie: I ate some of them.

However, on occasion *some* can be used in a context that does not exclude the total set. For example, in (4) Cookie Monster asserts that he has eaten some of the cookies but then goes on to explain that he ate all of them.

(4) Bert: If you ate some of the cookies, then I won't have enough for the party.
Cookie Monster: I ate some of the cookies. In fact, I ate all of them.

Grice argued these two interpretations are actually the result of a single meaning of *some* that is compatible with *all*. As Fig. 1 illustrates, the two terms, *some* and *all*, can be ordered on a scale with respect to the strength of the information that they convey (Horn, 1972, 1989; Gadzar, 1979). On this theory, the meaning of the weaker term (*some*) is consistent with all amounts greater than a lower boundary (*some* is greater than *none*) up through and including the maximum value (*all*). In sentences like (4), this meaning is transparent. Utterances with interpretations like this are termed lower-bounded since the scalar term has a lower boundary but no upper bound.

However, weaker scalar expressions are typically interpreted as having an upper boundary which excludes referents which are compatible with the maximal term (as in 3). This happens via a pragmatic inference called scalar implicature. According to Grice (1975), the participants in a conversation expect that each will tailor their contribution to be as informative as required but no more informative than is required (Quantity Maxim, pp. 45). Thus one can imagine a situation where Ernie had actually polished off the apples and uttered (5).

(5) Ernie: I ate all of the apples.

The existence of this more informative alternative means that if the speaker chooses instead to use a weaker scalar term like in (3), the listener can apply the Quantity Maxim and infer that this was a situation where the speaker is not in a position to make the stronger assertion (presumably because the stronger scalar term was not true). When this inference is made, the resulting interpretation is called upper-bounded since it imposes an additional boundary on the upper end of the scale. In other words, like Bert, we can infer that if Ernie had eaten all of the apples, he would have simply said so. Thus he must have eaten some-but-not-all of them. However, like all pragmatic inferences, the scalar implicature is defeasible allowing for the possibility of lower-bounded interpretations when the inference is cancelled or never calculated (as in 4).

This logic can be extended to any set of terms which can be placed on ordinal scale and which differ in their strength (Horn, 1972, 1989; Levinson, 2000). Parallel inferences have been noted for a wide range of expressions including scalar adjectives (warm vs. hot), aspectual verbs (start vs. finish), and logical operators (or vs. and). Thus if Ernie says he likes his soup *warm*, we can infer that he doesn't like it hot or if Bert says that he has *started* the book, we can infer that he hasn't finished it. Scalar implicatures can even be generated in cases where alternatives are ordered solely by virtue of the context or our knowledge of common practices (Hirschberg, 1985; Papafragou & Tantalou, 2004). For example, in sentence (1), our knowledge of making bread establishes a scale and as a result, the Hen's use of the weaker alternative (planted the grain) leads the listener to infer that the stronger is not true (made bread).

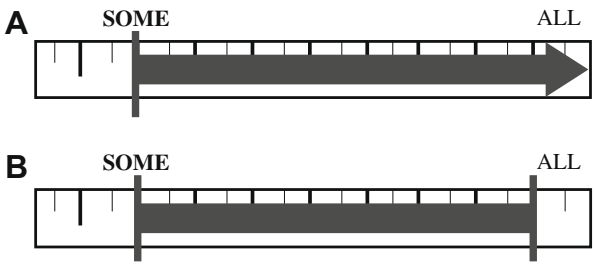

Fig. 1. Two interpretations of *some*. (A) Lower-bounded reading: the semantics of *some* can be described as referring to a ray of along a quantity scale. (B) Upper-bounded reading: *some* is typically interpreted with a pragmatic inference that excludes *all*.

In summary, the Gricean description of scalar implicature provides an explanation for the two readings of weak scalar terms which invokes constraints at two distinct levels of interpretation. At the semantic level, the meaning of *some* is always compatible with the total set (some-and-possibly-all). However at the pragmatic level, the interpretation can vary. Typically, an implicature will be calculated, as in (3), and *some* will be compatible with only a proper subset (some-but-not-all). However, this implicature is optional, and when it is absent or cancelled, as in (4), the pragmatic interpretation will have the same content as the semantic analysis. Note that the scalar implicature limits interpretation to a subset of the circumstances that are allowable on the basis of the semantic restriction alone. This creates an ideal situation for understanding the relationship between semantic and pragmatic processing since the facets of meaning that are assigned at each level analysis have consequences on the potential referents of the quantified phrase. In the remainder of the Introduction, we will first briefly review recent studies comparing semantic and pragmatic interpretation of various scalar terms using offline and online measures of language comprehension and then we will describe a series of experiments designed to isolate these components within real-time processing.

1.2. What can psycholinguistic studies tell us about scalar implicature?

Psycholinguistic studies have provided some empirical support for this two-level analysis of scalar interpretation. One source of evidence comes from research on developmental changes in the construal of scalar terms. These studies have demonstrated that while adults consistently favor upper-bounded readings, children prefer lower-bounded interpretations for a variety of scalar terms. For example, Noveck (2001) asked children and adults to evaluate statements like "x *might* be y" in contexts where "x *must* be y" was true. He found that while adults overwhelmingly rejected the weaker modal, seven- to nine-year-olds accepted it, suggesting that they treated the weaker statement as compatible with the stronger one. Similarly, Papafragou and Musolino (2003) found that five-year-olds, but not adults, were content to accept weak scalar predicates like *started* in situations where the stronger term *finished* applied.

In the absence of a distinction between semantics and pragmatics, this pattern would be puzzling. For example, if we assume that the scalars are simply semantically ambiguous (*some* means both some-and-possibly-all and some-but-not-all), then we confronted with a learning paradox. Adults typically use weak scalars in situations in which they are upper-bounded. Children, like adults, show a robust bias for the more common interpretation of an ambiguous word (Swinney & Prather, 1989). So how and why would children develop a preference for the less frequent lower-bounded readings? Noveck (2001) points out that the theoretical distinction between semantics and pragmatics allows us to make sense of this pattern: children, like adults, correctly retrieve a lower-bounded semantics for these scalar terms but, unlike adults, they fail to make the pragmatic implicature.

The nature of scalar implicature can also be explored by investigating the time-course of interpretation in adults. Several researchers have done this by comparing reaction times to sentences with implicatures to those without implicatures. For example, Bott and Noveck (2004) examined the response times for truth-value judgments of sentences containing weak scalar quantifiers like "Some elephants are mammals." For underinformative statements like these, participants' spontaneous judgments reveal how they are interpreting the sentence. False responses indicate an upper-bounded interpretation, while true responses indicate a lower-bounded one. They found that participants who judged the statements to be false took longer than those who judged them to be true. The authors attribute this difference to the time that it takes to generate the implicature. A similar data pattern has emerged in several other studies measuring speeded truth-value judgments of underinformative usages of *some* (De Neys & Schaeken, 2007; Posada & Noveck, 2003; Rips, 1975).

While these results are consistent with the two-level analysis described above, aspects of the method limit the conclusions we can draw. First, the use of underinformative statements introduces potential confounds. To link increases in reading time to one interpretation, the experimenters must either manipulate the participants' construal of the critical scalar or measure spontaneously occurring differences in the preferred analysis. If interpretation is directly manipulated (e.g., by instructing participants to analyze *some* as some-but-not-all or some-and-possibly-all), then we cannot be sure that the processes involved in deliberately carrying out this instruction are the same as those that would be

involved in ordinary comprehension. If we examine spontaneous variation in interpretation (e.g., by comparing trials where the pragmatic reading is accessed with those where the semantic reading prevails), then we necessarily move from an experimental design to a correlational one. This introduces the third-variable problem: the possibility that differences in reaction time between the two response types are attributable to some third factor which is responsible both for the longer reaction times and for the contrasting responses.

Feeney and colleagues have explored this latter possibility. They suggest that the reaction time differences between "pragmatic responders" and "logical responders" are attributable to differences in the participants' response strategies (Feeney, Scafton, Duckworth, & Handley, 2004). Thus they note that the participants in Noveck and Posada's study (2003) who select the upper-bounded response to the underinformative sentences also had slower reaction times to the other items, suggesting that these participants may simply be more cautious and systematic. They attribute the use of overt strategies to the large number of critical trials that were used and relative scarcity of filler items (raising awareness of the critical items). When the authors conducted a parallel study with fewer trials but more trial types, they found that individuals failed to adopt a consistent strategy and instead produced both "pragmatic" and "logical" responses. Critically, the within-subject comparison revealed no difference in the reaction times for the lower- and upper-bounded interpretations.

A second methodological limitation is that research on this topic has relied almost exclusively on sentence final judgments about the validity or truth of a statement. These judgments provide limited information about the processes that underlie the apparent delays. This is problematic for several reasons. First, since judgment times are not directly mapped onto separable periods of analyses, it is unclear whether scalar implicatures are ever preceded by a period of semantic analysis. Instead participants who spontaneously adopt an upper-bounded interpretation might do so in a single unitary process. Second, the use of verification tasks creates uncertainty about whether the increases in reaction times are actually attributable to linguistic processes, rather than processes involved in verification (Bott & Noveck, 2004). Is the delay caused by the time taken to calculate the implicature and thus restrict the meaning of the target utterance (the conversion of "Some elephants are mammals" to "Only some elephants are mammals")? Or does it reflect the need for additional time to test this more restrictive meaning against the participant's stored knowledge (the time required to ascertain that not only are there elephants that are mammals, but in fact there are no elephants which are not)?

These limitations are highlighted by differences in the findings across experiments. For example, while many investigators have found delays for upper-bounded interpretations, others have found no differences in the reaction times (Feeney et al., 2004) or delays for items in lower-bounded contexts (Bezuidenhout & Cutting, 2002).[1] If we take these increases in reaction times as indicative of additional stages in processing, these studies are difficult to reconcile. Feeney and colleagues (2004) have suggested that the pattern might be explained by a three stage process: first the logical reading is accessed, then in most contexts the implicature is calculated, and finally in some contexts this implicature can be cancelled restoring the lower-bounded reading. The direction of the reaction time difference will depend on whether the lower-bounded readings in a particular study reflect the first or third stage of processing. However, in the absence of detailed information about the time-course of interpretation this account remains speculative.

A recent paper by Breheny et al. (2006) addresses some of these limitations. The authors use a phrase-by-phrase self-paced reading task to examine the effects of context on the generation of implicatures for both *some* and *or*. This task has greater temporal resolution and places fewer demands on the participants. In the case of *or*, Breheny and colleagues found that scalar terms embedded in upper-bounded contexts were read more slowly than those in lower-bounded contexts, suggesting that the pragmatic inference involves an additional process which is not automatically triggered across all utterances (Experiment 1). In the case of *some*, additional support for this hypothesis is provided by the reading times for a continuation that presupposes the upper-bounded interpretation (Experiment

[1] These differences maybe specific to the stimuli used by the study. Breheny, Katsos, and Williams (2006) argue that many of the so-called upper-bounded items from Bezuidenhout and Cutting (2002) were not in fact genuine scalar implicatures but instead interpretations based on syntactic and semantic constraints.

3). Participants were presented with the upper-bounded or lower-bounded context seen in (6) and (7) and their reading times were compared during two critical regions following the quantifier.

(6) *Upper-bounded context:* Mary asked John whether he intended to host all his relatives in his tiny apartment. John replied that he intended to host *some of his relatives. The rest would stay* in a nearby hotel.

(7) *Lower-bounded context:* Mary was surprised to see John cleaning his apartment and she asked the reason why. John told her that he intended to host *some of his relatives. The rest would stay* in a nearby hotel.

They found that those who encountered the term in the upper-bounded context showed delays in reading the quantified phrase (“some of his relatives”), suggesting that the scalar implicature was calculated at this initial period. In contrast, those who encountered the term in the lower-bounded context demonstrated delays in the following region, in which the proper subset was explicitly referred to (“the rest would stay”), suggesting that the upper-bounded inference had not yet been made in the initial period.

While these studies are clearly consistent with the two-level analysis described above they also leave some questions open. First, because they rely on complex contextual manipulations to drive interpretation, it is difficult to pin the differences in reading times directly to the inference process. For example, in the study described above, the upper-bounded context not only emphasizes the need for a boundary, it also (1) contains considerably more overlap between the context sentence and the target sentence, (2) makes use of the contrasting scalar term (*all*), and (3) provides an antecedent in the discourse (“all his relatives”) for the critical scalar phrase (“some of his relatives”). It is not clear what the impact of each of these differences would be on reading times in the critical regions independent of their effects on implicature. Second, while these studies have greater temporal sensitivity, they still leave open the question of how participants arrive at the two analyses. Is the upper-bounded analysis slower because it is preceded by the lower-bounded one? Or does the delay merely reflect a difference in the length of a single process caused by a difference in complexity or accessibility of the two analyses?

1.3. Isolating the component processes

One way to circumvent these problems is to use a procedure that provides an indirect measure of comprehension as it takes place. The visual-world eye-tracking paradigm has been used extensively in psycholinguistic research to yield a sensitive, time-locked measure of linguistic processing (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Participants are presented with spoken instructions, asking them to manipulate objects within a visual reference world, while their eye movements to those objects are measured. This procedure has two advantages for exploring the relationship between semantics and pragmatics. First, since eye movements are typically made without conscious reflection, they provide a more implicit measure of comprehension prior to overt judgments, which may invoke higher-level strategic processes. Second, because eye movements are rapid, frequent and tightly linked to the processing of spoken language, they provide a fine-grained measure of how interpretation unfolds over time. Thus rather than having to infer the difficulty of a process based on the length of sentence final reaction times, these fixations provide information about the nature of the interpretation at a given point in time.

In the following experiments, we investigated how the processing of scalar terms unfolds over the course of online speech comprehension. Participants heard stories in which two types of objects were divided up between four characters, two boys and two girls. These stories were accompanied by a visual display. In the first experiment, the items were always divided such that one of the critical characters (e.g., the girls) had a proper subset of one item (e.g., the socks) while the other had the total set of second item (e.g., the soccer balls). In the critical condition, participants were given instructions like “Point to the girl that has some of the socks” (see Fig. 2) and their eye movements were recorded. These trials contained a period of semantic ambiguity beginning at the onset of the quantifier during which the referent of a lower-bounded reading of *some* is compatible with both of the critical characters.

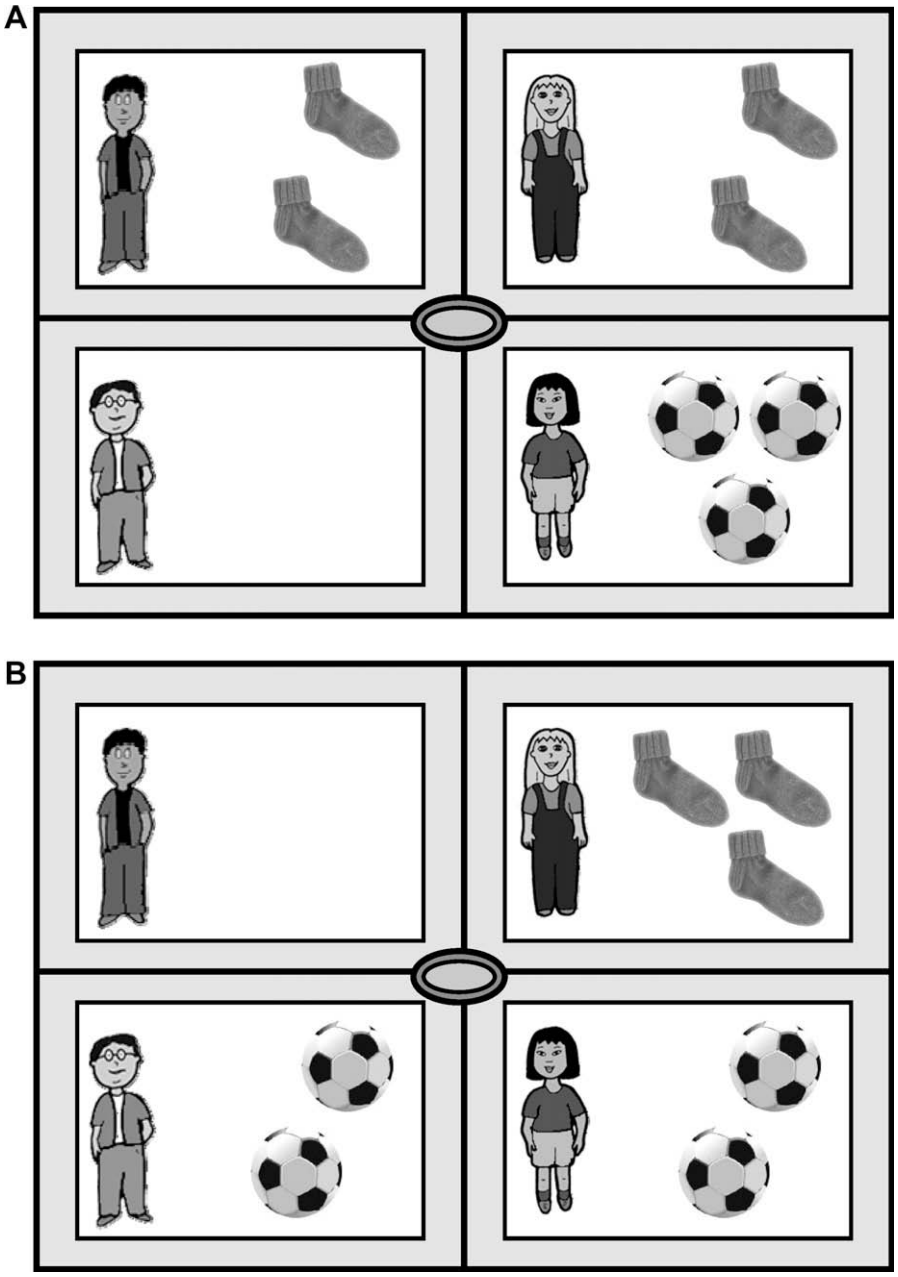

Fig. 2. In Experiment 1, examples of visual-world displays for (A) some/two trials and (B) all/ three trials. Participants here were instructed to "Point to the girl that has. . .of the socks." The girl with socks was the Target while the girl with soccer balls was the Distractor.

Eye movements to the target in this condition were compared to those in trials asking for "all of the socks" (in a context where one participant has all the socks and another has a proper subset of the soccer balls). In this case, the Distractor character (the girl with some-but-not-all of the soccer balls)

is inconsistent with the semantics of the quantifier. Thus if semantic meaning constrains interpretation prior to calculation of a pragmatic implicature, we would predict quick referential disambiguation in the *all* trials but prolonged competition between the two characters during the *some* trials. To ensure that differences between these trials were not simply due to preferences for larger quantities or a greater difficulty in calculating upper bounds relative to lower bounds, we also used terms from a number scale, *two* and *three*. Like *all*, these terms do not require a pragmatic inference to specify exact quantities and consequently do not have the same temporary semantic ambiguity as *some*.[2] Thus the performance on the *two* trials provides a crucial comparison since its meaning rules out the same competitor as *some* would once the implicature is calculated (the girl with soccer balls). By comparing these trials we can see whether there is any temporal delay between reference restriction via semantic content and reference restriction via pragmatic implicature.

2. Experiment 1

2.1. Methods

2.1.1. Participants

Twenty undergraduate students at Harvard University participated in this study. They received either course credit or $5 for their participation. All participants were native monolingual English speakers.

2.1.2. Procedure

Participants sat in front of an inclined podium divided into four quadrants, each containing a shelf where pictures could be placed (i.e. upper-left, upper-right, lower-left, and lower-right). A camera at the center of the display was focused on the participant's face and recorded the direction of their gaze while they were performing the task. A second camera, located behind the participant, recorded both their actions and the location of the items in the display. At the beginning of the study, the experimenter took out pictures of four characters and told the participant their names (i.e. Craig, Judy, Pat, and Cheryl). One picture was placed on each shelf in a pre-specified order and the participant was told that these boys and girls would receive different items throughout the experiment. Every trial consisted of a story context followed by a critical utterance. The experimenter acted out the stories using scripted utterances and pictures of the relevant objects. The story contexts were followed by the target utterances. These sound files were presented from an adjacent computer equipped with external speakers. Each target utterance instructed participants to pick up a particular character. Once the participant did this, the trial ended, the objects were removed from the display and the next trial began.

2.1.3. Materials

We can conceive of the four quantifiers as representing the four cells of 2×2 design in which the first factor, Quantifier Scale, contrasts terms derived from the critical Gricean scale (*some* and *all*) with terms from the control number scale (*two* and *three*). The second factor, Quantifier Strength, contrasts the position of these terms on an informational scale. For the Gricean scale, this factor distinguishes the weaker quantifier (*some*) from the stronger one (*all*) while for the number scale, it distinguishes the lesser quantity (*two*) from the greater one (*three*).

[2] This comparison between *some* and *two* is of interest for an additional reason since the semantics of number words has been an area of much contention within theoretical linguistics. The traditional Neo-Gricean account argues that numbers pattern with other scalar terms, possessing lower-bounded semantics (*two* means at least two) and receiving pragmatic upper bounds via scalar implicature (Gadzar, 1979; Horn, 1989; Levinson, 2000). More recently, others have argued that the semantics of number words do not pattern with other scalars and instead possess an exact semantics (Breheny, 2004; Horn, 1992; Koenig, 1991). While we have assumed for the purpose of these experiments that number words have lexically encoded upper bounds, we acknowledge that the time-course of the interpretation of these terms provides a simultaneous test of this hypothesis. That is, if number words have lower-bounded semantics, then we would predict the time-course of interpretation for *two* in this task should pattern with the interpretation of *some*. If, however, number words have exact semantics, we would expect *two* to pattern instead with *three* and *all* which are referentially unambiguous in this context.

The visual displays featured four characters that were aligned in the following clockwise order beginning from the upper-left quadrant: Craig, Judy, Cheryl, and Pat. This arrangement ensured that the vertically adjacent characters matched in gender while the horizontally adjacent characters did not (see Fig. 2). We constructed 16 stories like (8) below. In each story, two types of objects were introduced and distributed among the pairs of boys and girls.

(8) The boys and girls on the soccer team were getting socks and soccer balls from the coach. The coach gave socks to Judy and socks to Craig (*experimenter places two socks next to the girl on the upper-right and two socks next to the boy on the upper-left*). The coach knew that Pat was already a very good soccer player but he thought that Cheryl needed a lot of practice (*experimenter places a blank card next to the boy on the lower-left and three soccer balls next to the girl on the lower-right*).

In all cases, there was one set of four items that was split between a horizontally adjacent boy–girl pair and another set of three items which was given to one of the remaining children. By introducing the objects as part of a single large set and then dividing that set among the participants, we hoped to establish a frame of reference which would constrain the interpretation of the quantified phrases. For example, after the story given above, "all of the soccer balls" most naturally refers to all the soccer balls that the coach had, rather than all of the soccer balls in the known universe or all of the soccer balls that Cheryl has. In addition, the stories ensure that participants know the labels that we will be using for each object. In all the stories, the objects were referred to with definite noun phrases (e.g., "the socks") or bare plurals (e.g., "socks") to ensure that participants were not primed to associate a particular subset with the numbers and quantifiers used in the target utterances.

For each story we created a quartet of target sentences, like those shown in (9) below.

(9) Point to the girl that has *some/all/two/three* of the socks.

Because the story that preceded the target utterance was of a similar format regardless of the quantifier that was used, participants could not predict the quantifier based upon the configuration of the displays. The target sentences in each condition were identical except for the gender of the child that was requested and the identity of the final word. The gender of the child was linked to the content of the story: if the set of three objects had been given to a girl, then a girl was requested. The names of the two items that were distributed always had the same onset (e.g., socks and soccer balls), creating a brief period of ambiguity during which the identity of this noun was uncertain (see Appendix A for a list of all items).

Across all experiments, four versions of each base item were used to create four presentation lists such that each list contained four items in each condition and that each base item appeared just once in every list. No filler trials were included. There were three reasons why we considered them unnecessary: (1) the experiment was short with just four trials of each kind, minimizing the potential for the development of experimentally specific strategies; (2) the displays and stories were identical for *some* and *all* and for *two* and *three*, guaranteeing that the participant could not predict the quantifier prior to hearing it; (3) the use of both the quantifier scale and number scale ensured that each set could potentially be described in two ways, reducing the utility of strategic encoding.

The target sentences were recorded by a female actor. The digital waveforms were examined to ensure that the sentences had a consistent prosody, one which we thought was natural and unmarked. The sound files were edited to ensure that the lengths of two critical regions were equated across the four conditions: (1) the region from sentence onset to the gender cue ("Point to the") and (2) the region from the onset of the gender cue to the onset of the quantifier ("girl that has"). A trained research assistant coded the sentences using the ToBI annotation system (Beckman & Hirschberg, 1994). We verified the prosodic felicity of the target utterances by having a separate group of participants rate their naturalness with respect to the visual scene that was given. This data and the ToBI transcriptions are presented in Appendix B.

In this study, we relied upon the stories to create a context against which the utterances would be interpreted. It was critical that the participants interpreted phrases like "all of the soccer balls," as referring to all of the soccer balls in the display, not all of the soccer balls in the universe. Similarly, the logic of the experiment depended upon the participants' expecting that objects would be referred to using the basic-level terms that appeared in the story. If they expected phrases like

"some of the objects" or "all of the things," then eye movements at the onset of the quantifier would be uninformative. To verify that our contexts were successful in establishing the correct frame of reference, we administered a sentence completion task to see how participants interpreted the four quantifiers in the context of these displays. A separate group of 12 participants were tested on the 16 items from Experiment 1. Each participant heard the story and saw the display. Then they were given the target sentence with the final word removed (e.g., "Point to the girl that has some/all/two/three of the…"). Participants were told to fill in the blank in the way that made the most sense given the story and display. Items were rotated through lists as described above and each participant saw four items with each of the quantifiers. On every trial the participants completed the command with the label for the Target object which had been used in the story (e.g., *socks*). No participant used a superordinate term. If participants had a wider domain of quantification in mind, we might have expected them to provide a modifier for the *all* trials (e.g., "all of the socks that the coach had") but none of the participants did so. Furthermore, *some* was always taken to refer to the subset (*socks* rather than *soccer balls*) demonstrating that participants generated the implicature in an offline task.

2.1.4. Coding

Trained research assistants watched videotapes of the participants' actions and noted the character that was selected on each trial. Across all experiments, we only included trials where participants correctly selected the Target in subsequent analyses of eye movements. However, in Experiment 1 no trials were excluded on this basis. Approximately 0.9% of test trials were excluded from further analyses because of experimenter error.

Eye movements were coded by a research assistant who was blind to the location of each object, using frame-by-frame viewing of the participant's face on a Sony digital VCR. Each recorded trial began at the onset of the instruction and ended with completion of the corresponding action. Each change in direction of gaze was coded as towards one of the quadrants, at the center, or missing due to looks away from the display or blinking. These missing frames accounted for approximately 2% of all coded frames and were excluded from analysis. Afterwards the looks were then recoded based on their relation to the final instruction: (1) Target looks; (2) Distractor looks; (3) other looks to cards that did not match gender cues. The Target looks were defined as fixations to the character that matched both the gender and received item specified by the instruction (girl with socks) while the Distractor looks were fixations that matched the gender but not the received item (girl with soccer balls).

Twenty-five percent of the trials were checked by second coder who confirmed the direction of fixation for 93.6% of the coded frames. Any disagreements between the two coders were resolved by a third coder. For comparable displays, this method of analyzing eye movements has produced data similar to that collected using head-mounted eye-tracking (Snedeker & Trueswell, 2004).

2.2. Results

We examined the proportion of subjects' gaze time to the Target character on two different time scales. Our first analysis examined a coarse-grained measure of subjects' fixations as the target utterance unfolded. The second analysis focused in on fixations during the critical ambiguous region. The coarse-grained analysis examined five time windows:

(1) *Baseline phase:* This period begins at the onset of the instruction and ends just before the onset of the gender cue ("Point to the"). This region provides a baseline measure of looks to the display before the introduction of any gender or quantifier information. Here, we predict relatively equal fixations to the four characters.
(2) *Gender phase:* This period begins at the onset of the gender cue and ends just before the onset of the quantifier ("girl that has"). This region provides a direct comparison of looks to the Target and Distractor before the introduction of any quantifier information. Here, we predict that fixations would shift towards the side of the display with characters that match the specified gender.

(3) *Quantifier phase:* This critical period begins at the onset of the quantifier and ends just before the onset of the disambiguating phoneme ("some/all/two/three of the soc-"). Here, the comparison between the four trial types gives us a direct measure of any delay in reference restriction via semantic analysis vs. pragmatic inference. We predict that looks to the Target would rapidly increase relative to the Distractor when participants hear *all, two*, and *three*. Furthermore, if scalar implicatures are calculated immediately, we would also predict an early preference for the Target in the *some* trials. If, however, scalar implicatures are preceded by a period of initial semantic analysis, then we would expect looks to the Target and Distractor to remain relatively equal during this time.

(4) *Disambiguation phase:* This period begins at the onset of the disambiguating phoneme and ends at the offset of the command ("-ks"). This region unambiguously resolves the correct referent by picking out the unique item he or she possesses. Here, we predict fixations to be primarily focused on the Target with relatively few looks the Distractor.

(5) *End phase:* This period begins at the end of the sentence and continues for 900 ms. Again we predict that across all trials, participants would be looking at the Target prior to initiating their selection.

Table 1 lists the duration of the time windows. In our analyses, the onset of each time window is shifted 200 ms after the relevant marker in the speech stream to account for the time it would take to program a saccadic eye movement (Allopenna, Magnuson, & Tanenhaus, 1998; Matin, Shao, & Boff, 1993).

For all analyses, we use as our dependent measure total looking time to the Target as a proportion of looking time to the Target and the Distractor. This score ranged from zero (exclusive looks to the Distractor) to one (exclusive looks to the Target). Fixations to the other characters after onset of the gender cue were rare, accounting for less than 4% of total looks for all time points in the last three time windows. For this reason these looks were not included in the analyses. Each time window was analyzed with subjects and items ANOVAs with Quantifier Scale (number vs. scalar) and Quantifier Strength (lesser vs. greater) as within subject and item variables, and list/item group as a between subjects and items variable.

Fig. 3 illustrates that prior to the onset of the quantifier, the proportion of looks to the Target was around chance for all terms. Unsurprisingly, there were no reliable effects of Quantifier Scale or Strength during the Baseline and Gender phases (all F's < 1.00, all p's > .30). However during the Quantifier phase, fixations to the Target increased when participants heard *two* (68%), *three* (59%), and *all* (66%) but not when they heard *some* (48%). During this period, there was no main effect of Quantifier Scale ($F1(1,16) = 2.15$, $p > .10$; $F2(1,15) = 2.24$, $p > .10$) or Quantifier Strength ($F1(1,16) = 2.24$, $p > .10$; $F2(1,15) = 1.80$, $p > .10$). Critically, however, there was a significant interaction between the two variables ($F1(1,16) = 14.04$, $p < .01$; $F2(1,15) = 16.85$, $p < .01$). Planned comparisons within the levels of Quantifier Strength revealed that looks to the Target were significantly lower during the *some* trials than they were during the *two* trials ($t1(19) = 3.54$, $p < .01$; $t2(15) = 4.09$, $p < .01$) but there was no difference between the *all* and *three* trials ($t1(19) = 1.28$, $p > .20$; $t2(15) = 1.12$, $p > .20$). Comparisons within the Quantifier Scales revealed that fixations to the Target were greater during the *two* trials than they were during the *three* trials though this difference was only significant by subjects ($t1(19) = 2.10$, $p < .05$) but not by items ($t2(15) = 1.71$, $p > .10$). In contrast, the scalar quantifiers showed the opposite pattern: Target looks were significantly less in the *some* trials than they were in the *all* trials by both subjects and items ($t1(19) = 3.19$, $p < .01$; $t2(15) = 4.36$, $p < .01$). This demonstrates that participants were able to access the lexical semantics of the numbers and quantifiers and use these meanings to quickly disambiguate the referent. However, in the case of *some*, where

Table 1
In Experiments 1 and 2, duration of the time windows for the course-grained analysis

	Baseline phase	Gender phase	Quantifier phase	Disambiguation phase	End phase
Region length	600 ms	600 ms	800 ms	333 ms	900 ms

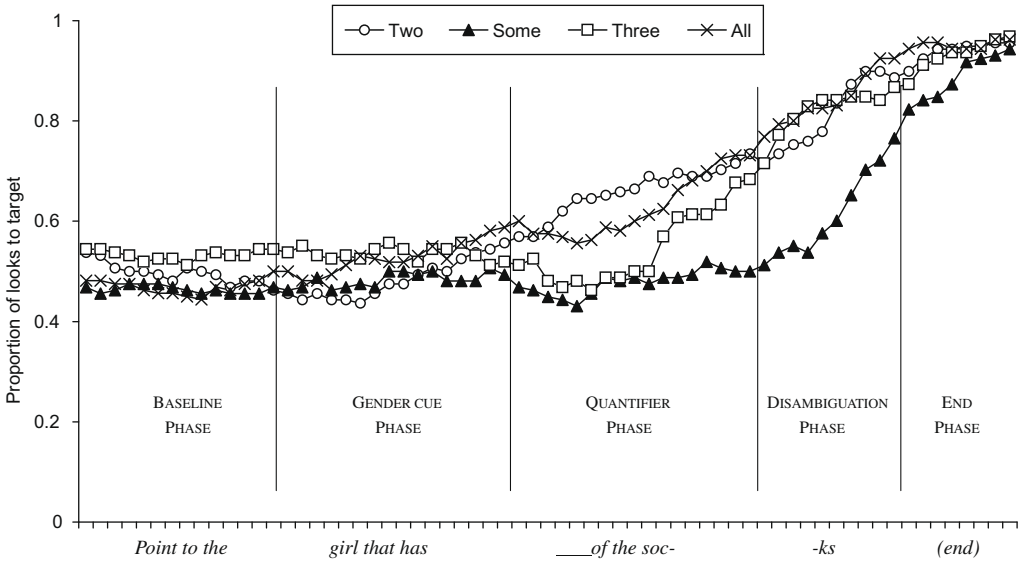

Fig. 3. In Experiment 1, the time-course of looks to Target for the four trial types.

the meaning of the quantifier was insufficient for disambiguation, there was a delay in reference resolution, suggesting that the pragmatic upper-bound was not available during this time.[3]

During the Disambiguation phase, looks to the Target character increased rapidly for the *some* trials (68%) suggesting that the phonological information allowed participants to close in on the correct referent even when the quantifier was referentially ambiguous (see Fig. 4). However, looks to the Target for the *some* trials continued to lag behind those in the *two* (86%), *three* (85%), and *all* (89%) trials. This difference lead to significant main effects of Quantifier Scale ($F1(1,16) = 8.61$, $p < .05$; $F2(1,15) = 3.28$, $p < .10$) and Quantifier Strength ($F1(1,16) = 12.39$, $p < .01$; $F2(1,15) = 3.70$ $p < .10$), and critically a significant interaction between these variables ($F1(1,16) = 12.55$, $p < .01$; $F2(1,15) = 11.54$, $p < .01$). Unsurprisingly, during the period immediately preceding the onset of the action, subjects in all conditions closed in on the Target. This lead to no differences across Scale ($F1(1,16) = 0.43$, $p > .10$; $F2(1,15) = .32$, $p > .10$) or Strength ($F1(1, 16) = 0.49$, $p > .10$; $F2(1,15) = 0.53$, $p > .10$), and no interaction between the two ($F1(1, 16) = 2.50$, $p > .10$; $F2(1,15) = 2.82$, $p > .10$).

The second analysis explored the critical interaction in greater detail. The proportion of fixations to the Target was calculated for 200 ms intervals beginning from the onset of the quantifier and continuing until but not including 1200 ms following quantifier onset. Each time window is defined by the period from the labeled time point to the frame prior to the onset of the next interval. Unlike the coarse-grain analyses, these intervals correspond to the real-time onset of speech information and were not shifted by 200 ms. Table 2 displays the proportion of looks to the Target for each quantifier type during each of these time windows. There was a significant Quantifier Scale by Strength interaction approximately 200 ms after the onset of the first phoneme of the quantifiers ($F1(1,16) = 8.29$, $p < .05$; $F2(1,15) = 7.69$, $p < .05$) which continued through the 1000 ms time window ($F1(1,16) = 29.67$, $p < .01$; $F2(1,15) = 13.85$, $p < .01$). During the initial time window, the proportion

[3] Careful readers (and an anonymous reviewer) have questioned whether we can attribute these differences across conditions to differences in the quantifiers, rather than differences in the displays. In our design the two are confounded: *some* and *two* refer to sets of two items while *all* and *three* refer to sets of three. Note however, that if the differences in the eye-movements were driven by the differences in the visual scenes, then we would expect looks in the *two* condition to pattern with looks in the *some* condition, but they do not. Instead they pattern with looks in the *three* and *all* conditions. Thus we ascribe these differences to the quantifiers. This conclusion is supported by the tight temporal relation between the quantifier onset and the emergence of the critical interaction.

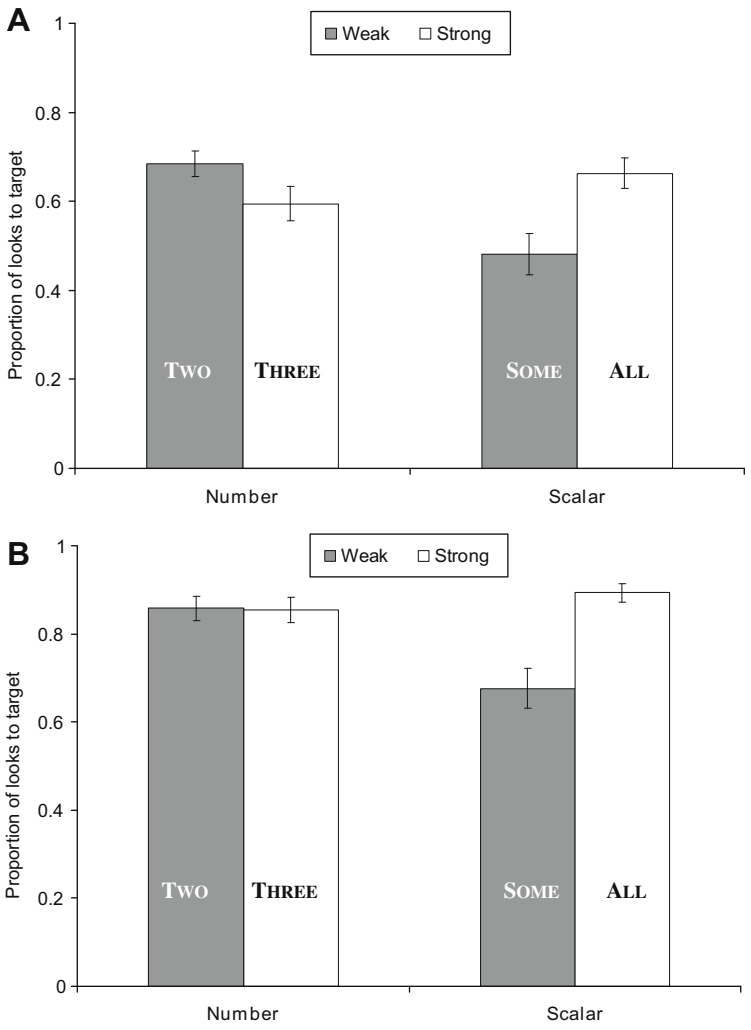

Fig. 4. In Experiment 1, the proportion of looks to Target during (A) the Quantifier phase and (B) the Disambiguation phase.

of looks to the Target during the *two* ($t1(19) = 2.91$, $p < .01$; $t2(15) = 2.51$, $p < .05$) and *all* ($t1(19) = 2.13$, $p < .05$; $t2(15) = 1.83$, $p < .10$) trials were significantly greater than chance. Target preference in *three* trials rose above chance in the 600 ms time window ($t1(19) = 3.05$, $p < .01$; $t2(15) = 6.77$, $p < .001$). In contrast, reference resolution in *some* trials was substantially delayed: Target preference was not significantly above chance until approximately 1000 ms after the onset of the quantifier ($t1(19) = 8.231$, $p < .001$; $t2(15) = 6.65$, $p < .001$). This time window overlaps with the phonological disambiguation of the critical noun, thus we see no evidence for referential disambiguation on the basis of a scalar implicature.

2.3. Discussion

In Experiment 1, we found that participants' reference resolution was strongly affected by the term they heard. In our coarse-grained analysis of fixations, there was an increase in looks to the Target dur-

Table 2
In Experiment 1, the proportion of looks to the target for each quantifier type during each 200 ms time window

	Time from onset of the quantifier (in ms)					
	0	200	400	600	800	1000
Two	0.53 ($CI_{\pm 95}$ = .03)	0.61 ($CI_{\pm 95}$ = .04)	0.68 ($CI_{\pm 95}$ = .03)	0.72 ($CI_{\pm 95}$ = .03)	0.82 ($CI_{\pm 95}$ = .03)	0.93 ($CI_{\pm 95}$ = .03)
Some	0.50 ($CI_{\pm 95}$ = .04)	0.47 ($CI_{\pm 95}$ = .04)	0.47 ($CI_{\pm 95}$ = .05)	0.50 ($CI_{\pm 95}$ = .06)	0.57 ($CI_{\pm 95}$ = .05)	0.79 ($CI_{\pm 95}$ = .04)
Three	0.54 ($CI_{\pm 95}$ = .04)	0.49 ($CI_{\pm 95}$ = .04)	0.55 ($CI_{\pm 95}$ = .05)	0.66 ($CI_{\pm 95}$ = .05)	0.85 ($CI_{\pm 95}$ = .03)	0.89 ($CI_{\pm 95}$ = .03)
All	0.56 ($CI_{\pm 95}$ = .04)	0.59 ($CI_{\pm 95}$ = .04)	0.64 ($CI_{\pm 95}$ = .04)	0.72 ($CI_{\pm 95}$ = .04)	0.83 ($CI_{\pm 95}$ = .03)	0.96 ($CI_{\pm 95}$ = .02)
ANOVA (strength by scale)	$F1(1,16) = 0.48$, $p > .40$; $F2(1,15) = 0.43$, $p > .50$	$F1(1,16) = 8.29$, $p < .05$; $F2(1,15) = 7.69$, $p < .05$	$F1(1,16) = 16.20$, $p < .01$; $F2(1,15) = 17.54$, $p < .01$	$F1(1,16) = 10.04$, $p < .01$; $F2(1,15) = 13.85$, $p < .01$	$F1(1,16) = 8.17$, $p < .05$; $F2(1,15) = 10.74$, $p < .01$	$F1(1,16) = 29.67$, $p < .001$; $F2(1,15) = 13.85$, $p < .01$

ing the Quantifier phase for *two, three*, and *all* trials but not for the *some* trials. Thus when lexical semantics is sufficient to identity the Target, disambiguation is quite rapid. When semantic analysis is not sufficient, as in the case of *some*, reference resolution is substantially delayed. In fact, under these circumstances participants fail to show a reliable Target preference until the Disambiguation region, suggesting that rather than calculating the pragmatic inference they simply wait until the final phoneme of the phrase indicates the correct referent (i.e. use *-ks* to select *socks* rather than *soccer balls*). This delay is particularly remarkable since in this experiment participants were only given utterances and displays that were consistent with the preferred some-but-not-all interpretation. Thus the delay in interpretation occurs even in contexts where the implicature is never violated and the referent is ultimately unambiguous.[4]

Finally, our fine-grained analyses produced one unexpected finding. While a Target preference developed rapidly after quantifier onset in the *two* and *all* trials, it was delayed by an additional 400 ms in the *three* trials. Since *all* and *three* picked out the same referent in this paradigm, this delay cannot be attributed to perceptual features of the Target. Similarly, since *two* and *three* belong to the same scale and have parallel semantics by all accounts, this slowness is unlikely to reflect differences in meaning. One possible explanation for this peculiarity lies in the interaction between the linguistic structure of our commands and the arrangement of our displays. Recall that participants were always asked to select the character with "...of the socks." This partitive construction highlights the relationship between subsets and total sets (Chierchia, 2004b) and we employed it in this task because it unambiguously generates a scalar implicature when used with *some*. This construction is also felicitous for the *two* trials, in which a set of two items is selected from a total set of four, and for the *all* trials in which the quantifier unambiguously picks out the total set. However, this construction is infelicitous for the *three* trials which request the person with "three of the socks" in a context in which there are only three socks. Thus delays in these trials might have reflected the awkward nature of

[4] An anonymous reviewer raised the question of whether participants might learn to encode the subset as *some* over the course of the study and as a result, show less of a delay in resolving the Target. We explored this possibility by conducting an ANOVA comparing the first and second half of the experiment. We found no effects of half or interactions between half and the critical variables (all *p*'s > .15). Looks to the Target in the *some* trials lagged behind those in the *two*, *three*, and *all* trials in both halves of the study. Two features of the experiment may have discouraged participants from strategically encoding the stimuli in this way. First, there were relatively few critical items, thus while *some* was always paired with the subset, this pairing occurred only four times in the experiment. Second, by including both the number and scalar trials, we ensured that each character could be described in two ways (e.g., the girl with two of four socks can be construed either as *two* or *some* of the socks), making it impossible to predict precisely which linguistic label would be used. These features allowed us to more accurately assess participants' biases based on prior experiences with these quantifiers rather than specific mappings established within the experimental context.

using a partitive construction in situation in which the items in question are not readily construed as a subset of some other set.

In Experiment 2, we sought to replicate our findings while simultaneously addressing this concern. To ensure that the partitive construction would be felicitous for both *two* and *three*, we modified the displays for the number trials by changing the number of objects given to the character of the opposite gender. For the *three* trials, the boy who previously received no socks now received one sock and for the *two* trials the boy who previously received no soccer balls now received one soccer ball. The presence of these additional objects should permit the felicitous use of the partitive construction for the *three* trials since the subset of three is now part of a larger total set. The configurations for the quantifier trials remained as they were in Experiment 1. Consequently the distribution of the objects could potentially alert the participants to the kind of quantifier that would be used, but it could not alert them to the strength of the quantifier. Furthermore, since the additional object in the number trials was always assigned to a character that was ruled out by the gender cue, the Target and Distractor characters continued to be identical for all four trial types. Thus, we expect that this additional object will not have a direct influence on looks to the critical items following the onset of the quantifier.

3. Experiment 2

3.1. Methods

3.1.1. Participants

Twenty undergraduate students at Harvard University participated in this study. They received either course credit or $5 for their participation. All were native monolingual English speakers who had no history of participation in the previous experiment. Two additional students took part in the study but were excluded from these analyses due to experimental error.

3.1.2. Procedure

The procedure was identical to Experiment 1.

3.1.3. Materials

The materials were similar to Experiment 1 with one key difference: the distribution of objects among the four characters differed in the two Quantifier Scale conditions. For the *some* and *all* trials, the arrangement remained unchanged—one set of four items was split evenly between a horizontally adjacent pair (girl with two socks and boy with two socks) and another set of three items which was given to one child from the remaining pair (girl with three soccer balls and boy with no soccer balls). For the *two* and *three* trials, the first set was again evenly split between one boy–girl pair while the second set now included a fourth item given to the character who had previously received nothing (boy with one soccer balls).

3.1.4. Coding

The data was coded in the manner described in Experiment 1. Approximately 0.9% of trials were excluded from further analysis due to experimenter error while approximately 0.3% of trials were excluded because of a participant's incorrect action. Finally, missing frames due to blinks or looks away accounted for 3% of all coded frames and were also excluded from analysis. First and second coding had 95.1% inter-coder reliability.

3.2. Results

3.2.1. Analyses of proportion of looking time to the Target

We again examined the proportion of subjects' looking time to the Target as a proportion of looking time to the Target and Distractor characters, using the same coarse- and fine-grained time windows employed in Experiment 1 (see Fig. 5). Fixations to the other characters after onset of the gender cue were rare, accounting for less than 3% of total looks in the Quantifier, Disambiguation, and End phases.

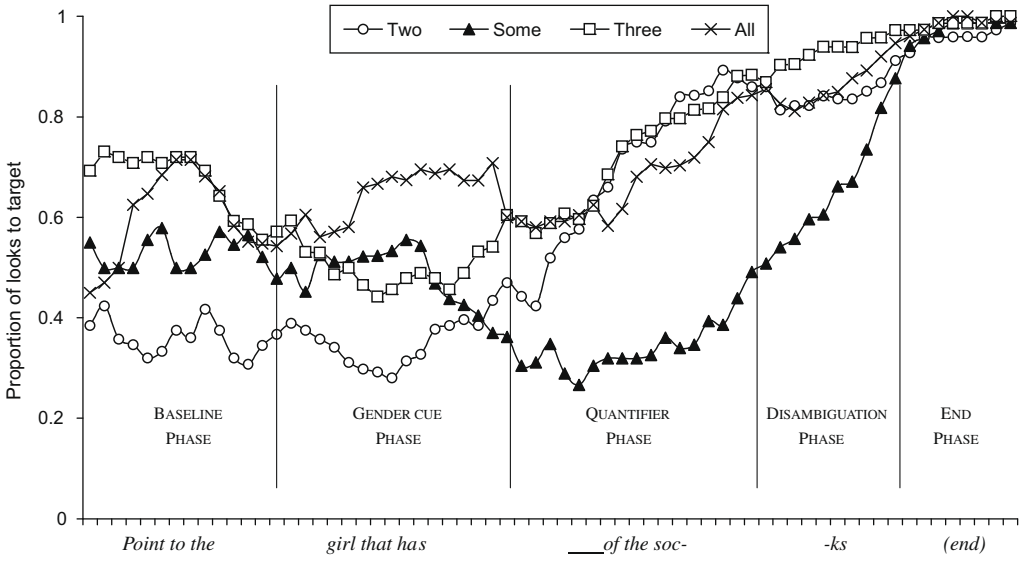

Fig. 5. In Experiment 2, the time-course of looks to Target for the four trial types.

In the coarse-gained time windows, there were no reliable effects of Quantifier Scale or Strength during the Baseline and Gender phases (all F's < 4.00, all p's > .05). This changed during the Quantifier phase where fixations to the Target character increased when participants heard *two* (66%), *three* (72%), and *all* (72%) but not when they heard *some* (45%). During this period, there were both main effects of Quantifier Scale ($F1(1,16) = 5.16$, $p < .05$; $F2(1,15) = 6.39$, $p < .05$) and Quantifier Strength ($F1(1,16) = 16.86$, $p < .01$; $F2(1,15) = 18.29$, $p < .01$). Critically, there was also the predicted significant interaction between these variables ($F1(1,16) = 6.58$, $p < .05$; $F2(1,15) = 5.25$, $p < .05$). Planned comparisons within the levels of Quantifier Strength revealed that looks to the Target in the *some* trials were significantly lower than in the *two* trials ($t1(19) = 3.22$, $p < .01$; $t2(15) = 3.11$, $p < .01$) but there was no reliable difference between the *all* and *three* trials ($t1(19) = 0.01$, $p > .50$; $t2(15) = 0.04$, $p > .50$). Comparisons within the Quantifier Scales revealed that there was no difference between *two* and *three* trials ($t1(19) = 1.08$, $p > .20$; $t2(15) = 0.98$, $p > .30$) but a reliable difference between the *some* and *all* trials ($t1(19) = 3.93$, $p < .01$; $t2(15) = 4.15$, $p < .01$).

However, unlike Experiment 1, this pattern quickly disappeared after the onset of the final phoneme (see Fig. 6). Fixations to the Target character during the Disambiguation phase increased for all trial types (82% for the *two* trials, 91% for the *three* trials, 86% for the *all* trials, and 71% for the *some* trials). In this region there was a significant effect of Quantifier Strength ($F1(1,16) = 15.65$, $p < .01$; $F2(1,15) = 23.66$, $p < .01$) but no effect of Quantifier Scale ($F1(1,16) = 3.19$ $p > .05$; $F2(1,15) = 3.20$, $p > .05$) and no interaction ($F1(1,16) = 0.73$, $p > .10$; $F2(1,15) = 0.63$, $p > .10$). Finally, during the End phase, total fixations closed in unsurprisingly on the Target leading to no effect of Quantifier Scale ($F1(1,16) = 0.24$, $p > .10$; $F2(1,15) = 0.07$, $p > .10$) and Strength ($F1(1,16) = 0.78$, $p > .10$; $F2(1,15) = 0.48$, $p > .10$), and no interaction between them ($F1(1,16) = 0.32$, $p > .10$; $F2(1,15) = 0.29$, $p > .10$).

Additional analyses of 200 ms intervals following the quantifier onset confirmed the difference in time it took participants to reliably fixate on the Target character across the four terms. Table 3 displays the proportion of looks to the Target for each quantifier type during each of these time windows. There was a significant Quantifier Scale by Strength interaction that began approximately 400 ms after the onset of the quantifier ($F1(1,16) = 6.78$, $p < .05$; $F2(1,15) = 7.18$, $p < .05$) and continued into the 600 ms time window ($F1(1,16) = 11.09$, $p < .01$; $F2(1,15) = 14.19$, $p < .01$). During this period, the proportion of looks to the Target on the *two* ($t1(19) = 4.77$, $p < .001$; $t2(15) = 3.62$, $p < .01$), *three* ($t1(19) = 4.20$, $p < .001$; $t2(15) = 4.76$, $p < .001$), and *all* ($t1(19) = 2.82$, $p < .05$; $t2(15) = 3.22$, $p < .01$) tri-

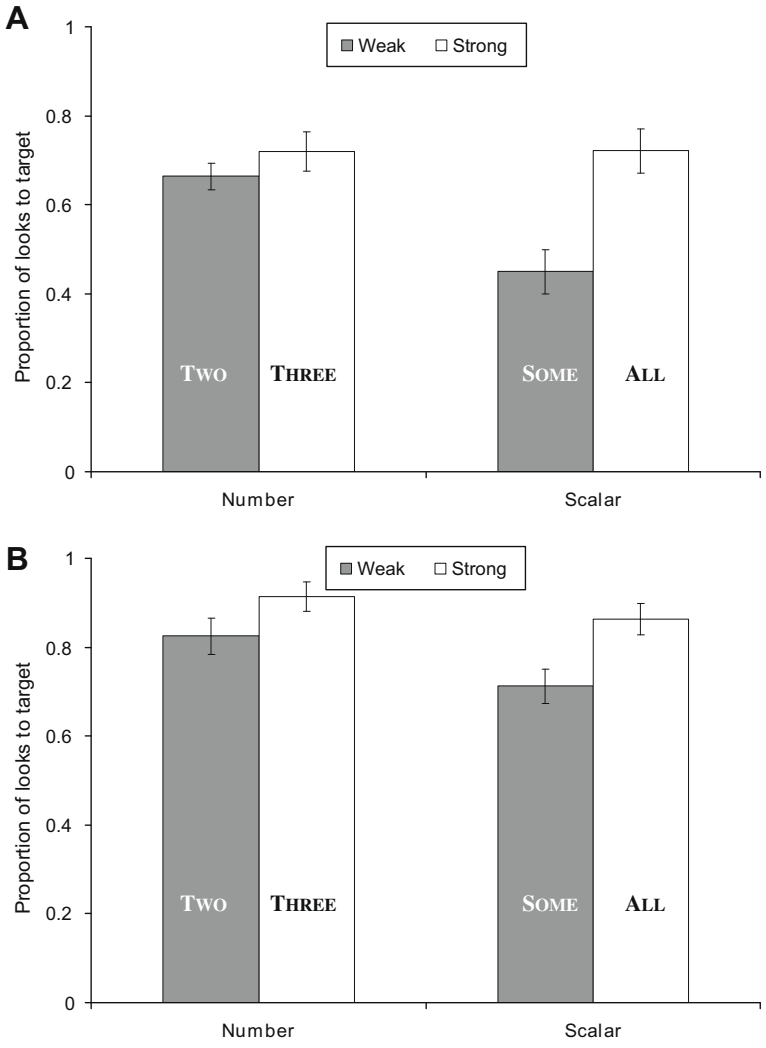

Fig. 6. In Experiment 2, the proportion of looks to Target during (A) the Quantifier phase and (B) the Disambiguation phase.

als were significantly greater than chance. In contrast, the preference for the Target in the *some* trials emerged later and was not significantly above chance until approximately 800 ms following the onset of the quantifier ($t1(19) = 2.24$, $p < .05$; $t2(15) = 1.99$, $p < .10$).

3.2.2. Analyses of eye movements initiated after the quantifier

While we found no significant differences across conditions before the onset of the quantifier, visual inspection of Fig. 5 suggests that there was a tendency for participants to look at characters who had more objects. During the Baseline and Gender phases, participants in the *some* and *two* trials tended to look at the Distractor while those in the *all* and *three* trials tended to look at the Target. This raises the possibility that the delay in Target fixations during the Quantifier region for *some* may simply reflect the continuation of this bias to look at larger quantities.

There are two reasons why we do not believe this to be the case. First, we see the same perceptual bias in the number trials prior to the onset of the quantifier, where looks to the Target is greater for

Table 3
In Experiment 2, the proportion of looks to the target for each quantifier type during each 200 ms time window

	Time from onset of the quantifier (in ms)					
	0	200	400	600	800	1000
Two	0.43 ($CI_{\pm 95}$ = .03)	0.53 ($CI_{\pm 95}$ = .04)	0.67 ($CI_{\pm 95}$ = .04)	0.79 ($CI_{\pm 95}$ = .04)	0.79 ($CI_{\pm 95}$ = .05)	0.88 ($CI_{\pm 95}$ = .03)
Some	0.45 ($CI_{\pm 95}$ = .05)	0.39 ($CI_{\pm 95}$ = .04)	0.41 ($CI_{\pm 95}$ = .05)	0.48 ($CI_{\pm 95}$ = .06)	0.61 ($CI_{\pm 95}$ = .05)	0.84 ($CI_{\pm 95}$ = .03)
Three	0.51 ($CI_{\pm 95}$ = .05)	0.57 ($CI_{\pm 95}$ = .05)	0.70 ($CI_{\pm 95}$ = .05)	0.80 ($CI_{\pm 95}$ = .04)	0.86 ($CI_{\pm 95}$ = .04)	0.95 ($CI_{\pm 95}$ = .02)
All	0.60 ($CI_{\pm 95}$ = .04)	0.57 ($CI_{\pm 95}$ = .04)	0.64 ($CI_{\pm 95}$ = .05)	0.78 ($CI_{\pm 95}$ = .05)	0.82 ($CI_{\pm 95}$ = .04)	0.93 ($CI_{\pm 95}$ = .02)
ANOVA (strength by scale)	$F1(1,16) = 0.93$, $p > .30$; $F2(1,15) = 1.18$, $p > .29$	$F1(1,16) = 2.47$, $p > .10$; $F2(1,15) = 3.35$, $p > .05$	$F1(1,16) = 6.78$, $p < .05$; $F2(1,15) = 7.18$ $p < .05$	$F1(1,16) = 11.09$, $p < .01$; $F2(1,15) = 14.19$, $p < .01$	$F1(1,16) = 3.03$, $p > .10$; $F2(1,15) = 2.74$, $p > .10$	$F1(1,16) = 0.13$, $p > .70$; $F2(1,15) = 0.20$, $p > .60$

three compared to *two*. However, this difference disappears immediately after the quantifier is heard, suggesting that fixations during the critical regions are driven primarily by responses to the linguistic input. Second, in the absence of any such differences in Experiment 1, we still found slower Target fixations in the *some* trials. This suggests that delays in reference resolution for *some* cannot be fully explained by prior perceptual bias.

Nevertheless, we conducted an additional analysis of fixations initiated after quantifier onset, so we could determine whether the differences between *some* and other quantifiers persisted when these baseline differences are factored out. To do this, we divided the trials based on what the participant was fixating on during the frame immediately preceding the onset of the quantifier. On trials where they were initially fixated on one of the Non-Target characters, we calculated the probability that they had switched their gaze to the Target for time windows following the onset of the quantifier. On trials where they were initially fixated on the Target, we calculated the probability that they abandoned the correct referent in favor of another character. Analyses of this kind have been used extensively in research on the development of word recognition (Fernald, Pinto, Swingley, Weinberg, & McRoberts, 1998; Swingley & Fernald, 2002; see also Altmann & Kamide, 2004) and allow us to factor out early differences in fixation patterns by specifically comparing trials on which participants were looking at the same objects when the quantifier began. If participants' slowness in *some* trials (relatively to *two, three*, and *all* trials) were solely driven by the preferences in the region before the quantifier, then we would expect these differences to disappear in the present analyses. If, however, participants' fixations reflect a delay in calculating the implicature, then we would again expect slower latency to the Target following *some* compared to *all, three*, and *two*.

For each of these subsets of data we analyzed looking time in 100 ms intervals from the onset of the quantifier until 1000 ms after quantifier onset (Fig. 7). Each time window was analyzed with two-way ANOVAs with Quantifier Scale (number vs. scalar) and Quantifier Strength (lesser vs. greater) as within subject and item variables, and list/item group as a between subjects and items variable. Table 4 illustrates that on the Non-Target initial trials, the pattern of fixations across the four trial types began to differ about 600 ms after the onset of the quantifiers ($F1(1,16) = 8.16$, $p < .05$; $F2(1,15) = 12.85$, $p < .01$) and continued until 800 ms after quantifier onset ($F1(1,16) = 4.51$, $p < .05$; $F2(1,15) = 1.90$, $p > .15$). Following the onset of *two, three*, and *all*, participants began switching to the Target item, presumably because the semantics of these terms ruled out the Non-Target character (typically the Distractor). Planned comparisons revealed that there were no differences in switches to the Target between the number words (*two* vs. *three*) or between the strong terms (*three* vs. *all*, all *p*'s > .15). In contrast, following the onset of *some*, participants continued looking at a Non-Target character, suggesting that they had initially failed to calculate the scalar implicature. As a result, the proportions of switches

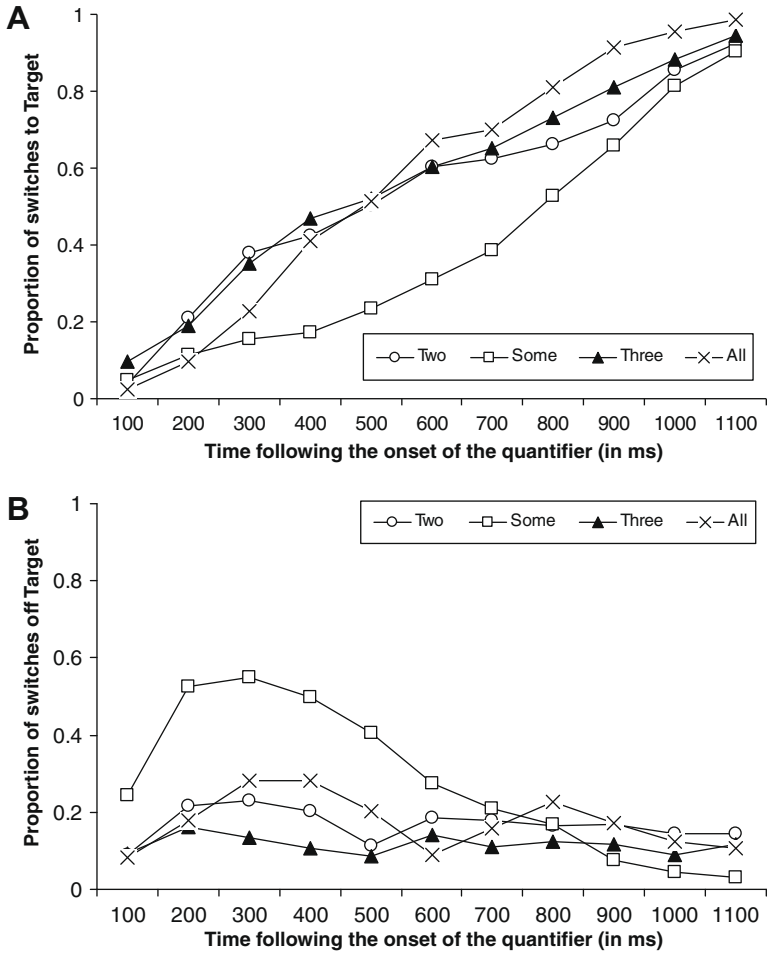

Fig. 7. In Experiment 2, trials were separated based on fixations prior to the onset of the quantifier (A) Non-Target initial trials: the proportion of switches to the Target and (B) Target initial trials: the proportion of switches off the Target.

to the Target were significantly lower during the *some* trials than during the *all, three*, and *two* trials (all p's < .05).

A similar pattern emerged in the Target initial trials (Table 5). The pattern of fixations across the four trial types began to differ about 200 ms after the onset of the quantifiers ($F1(1,16) = 4.59$, $p < .05$; $F2(1,15) = 7.73$, $p < .05$). Following the onset of *two, three*, and *all*, participants adhered to their initial looks to the Target, presumably because the Distractor was inconsistent with the semantics of these terms. Planned comparisons again revealed no differences in switches to the Target for the number words (*two* vs. *three*) or for the strong terms (*three* vs. *all*, all p's > .15). In contrast, following the onset of *some*, participants often abandoned their initial looks to the Target, suggesting that a scalar implicature was not initially available to restrict looks to the subset. As a result, the proportions of switches off the Target were significantly higher during the *some* trials than during the *all, three*, and *two* trials (all p's < .05).

Thus the analyses of both the Target initial trials and the Non-Target initial trials confirm that while listeners rapidly use the meanings of *all, three* and *two* to restrict reference, they do not use the pragmatic upper bound of *some* to guide their initial interpretation of the utterance.

3.2.3. Was the scalar implicature calculated online?

In the critical *some* condition, participants showed a reliable preference for the Target in the 800 ms time window. Since this time window preceded the disambiguating phoneme, we can conclude that participants were using some other source of information to identify the Target. There are two possibilities. First, participants could be generating the implicature online, just prior to the disambiguating phoneme. Second, participants could be using acoustic cues in the ambiguous region to determine whether the noun will refer to the Target item (socks) or the Distractor item (soccer ball). While we spliced the audio to equate instructions prior to the onset of the quantifiers, we did not splice the audio in the ambiguous region. Two acoustic cues could potentially disambiguate the utterance. Because some of the Target and Distractor pairs had the same syllable onsets but different codas, co-articulatory information on the vowel could give away the identity of the noun. In addition some of the Target and Distractor items differed in the number of syllables that they contained. Since utterance final syllables are typically longer than those in the middle of an utterance, the length of the vowel in the ambiguous onset may have provided participants with information about the length of the final noun.

To explore this question in more depth we examined the subset of trials in which the Target and Distractor object had the same number of syllables (eight of the 16 items). To rule out possible effects of co-articulatory information on the vowel, we identified the earliest point at which the vowel began (800 ms after the quantifier onset) and focused our analysis on the time region immediately preceding this (600–800 ms). Our goal was to ascertain whether eye movements executed during this interval showed a preference for the Target. To do this we split the *some* trials based on the object that the participant was fixating in the previous frame (Target or Distractor) and calculated the probability of switching to the other object during this interval. We found that participants were more likely to switch to the Target on the Distractor initial trials, than they were to switch to the Distractor on Target initial trials (33% vs. 7%; $t1(19) = 3.06$, $p < .01$; $t2(7) = 5.86$, $p < .01$). Since these switches occurred before the vowel, and these trials provided no prosodic cues to Target identity, we conclude that this Target preference suggests that the implicature is calculated by 800 ms after quantifier onset on some portion of the trials.

While this study provides some preliminary evidence that scalar implicature are calculated online, this data clearly demonstrates that there is a temporal lag between semantic and pragmatic influences on reference resolution in this task. In a parallel analyses of the time windows from 400 to 600 ms after quantifier onset, we found a robust Target preference in shifts for both *two* and *all* ($t1(19) = 3.35$, $p < .01$; $t2(7) = 1.54$, $p > .15$ and $t1(19) = 2.10$, $p < .05$; $t2(7) = 1.83$, $p > .10$, respectively), but not for *some* ($t1(19) = 1.70$, $p > .10$; $t2(7) = 0.08$, $p > .90$). Thus while the lexical semantics of the quantifier can be used to disambiguate reference by 400 ms of quantifier onset, scalar implicature fails to have an effect until 400 ms later.

3.3. Discussion

In Experiment 2, we replicated the interaction between Quantifier Scale and Strength in the proportion of fixations to the Target. During the Quantifier phase, we found increased fixations to the Target for the *two, three*, and *all* trials suggesting that listeners were able to quickly use the lexical semantics of the number words and the strong scalar quantifier to disambiguate the referent. Once again we found a delayed disambiguation for the *some* trials, suggesting that initial processing was limited to the lower-bounded lexical semantics of this weak quantifier. This apparent failure to immediately generate a pragmatic implicature was confirmed in the fine-grained analyses where a reliable preference for the Target emerged 400 ms after the quantifier for the *two, three*, and *all* trials but did not appear until the 800 ms time window for the *some* trials. Finally, we demonstrated the same asymmetry appeared in analyses of fixations occurring after the onset of the quantifier. The absence of an immediate implicature led to both fewer switches to the correct Target and more switches to an incorrect Non-Target in the *some* trials.

There were two key differences between Experiments 1 and 2. First, we found that the significant Scale by Strength interaction was limited to the period of semantic ambiguity in the Quantifier phase and disappeared during the Disambiguation phase. This difference is largely driven by an increase in

Table 4
In Experiment 2, the probability of switching looks to the Target in the Non-Target initial trials for each quantifier type during each 100 ms time window

	Time from onset of the quantifier (in ms)								
	200	300	400	500	600	700	800	900	1000
Two	0.21 ($CI_{\pm 95}$ = .05)	0.38 ($CI_{\pm 95}$ = .06)	0.42 ($CI_{\pm 95}$ = .06)	0.50 ($CI_{\pm 95}$ = .08)	0.60 ($CI_{\pm 95}$ = .08)	0.62 ($CI_{\pm 95}$ = .08)	0.66 ($CI_{\pm 95}$ = .08)	0.72 ($CI_{\pm 95}$ = .08)	0.85 ($CI_{\pm 95}$ = .05)
Some	0.11 ($CI_{\pm 95}$ = .04)	0.15 ($CI_{\pm 95}$ = .04)	0.17 ($CI_{\pm 95}$ = .05)	0.23 ($CI_{\pm 95}$ = .06)	0.31 ($CI_{\pm 95}$ = .07)	0.38 ($CI_{\pm 95}$ = .07)	0.53 ($CI_{\pm 95}$ = .07)	0.66 ($CI_{\pm 95}$ = .06)	0.82 ($CI_{\pm 95}$ = .04)
Three	0.19 ($CI_{\pm 95}$ = .07)	0.35 ($CI_{\pm 95}$ = .08)	0.47 ($CI_{\pm 95}$ = .08)	0.52 ($CI_{\pm 95}$ = .07)	0.60 ($CI_{\pm 95}$ = .07)	0.65 ($CI_{\pm 95}$ = .07)	0.73 ($CI_{\pm 95}$ = .07)	0.81 ($CI_{\pm 95}$ = .06)	0.88 ($CI_{\pm 95}$ = .04)
All	0.10 ($CI_{\pm 95}$ = .06)	0.23 ($CI_{\pm 95}$ = .07)	0.41 ($CI_{\pm 95}$ = .09)	0.51 ($CI_{\pm 95}$ = .08)	0.67 ($CI_{\pm 95}$ = .09)	0.70 ($CI_{\pm 95}$ = .08)	0.81 ($CI_{\pm 95}$ = .06)	0.91 ($CI_{\pm 95}$ = .03)	0.96 ($CI_{\pm 95}$ = .02)
ANOVA (strength by scale)	$F1(1,16) = 0.01$, $p > .90$; $F2(1,15) = 0.24$, $p > .60$	$F1(1,16) = 1.38$, $p > .20$; $F2(1,15) = 1.53$, $p > .20$	$F1(1,16) = 2.95$, $p > .10$; $F2(1,15) = 1.84$ $p > .15$	$F1(1,16) = 4.24$, $p < .10$; $F2(1,15) = 5.59$, $p < .05$	$F1(1,16) = 8.16$, $p < .05$; $F2(1,15) = 12.85$, $p < .01$	$F1(1,16) = 6.20$, $p < .05$; $F2(1,15) = 5.15$, $p < .05$	$F1(1,16) = 6.44$, $p < .05$; $F2(1,15) = 2.99$, $p > .10$	$F1(1,16) = 4.51$, $p < .05$; $F2(1,15) = 1.90$, $p > .15$	$F1(1,16) = 2.14$, $p > .15$; $F2(1,15) = 1.51$, $p > .20$

Target fixations following the onset of *three* in Experiment 2 as compared to Experiment 1 (72% vs. 59%, respectively, $t1(38) = 2.13$, $p < .05$; $t2(30) = 2.25$, $p < .05$). This suggests that with the addition of the extra-Distractor object, *three* was no longer infelicitous in the partitive construction. Second, the two experiments also differed in the timing of shifts to the Target for the critical *some* condition. A comparison of the fine-grained analyses demonstrates that in Experiment 2 but not in Experiment 1, the proportion of looks to the Target was reliably above chance in the time window which began 800 ms after the onset of the quantifier. This time window preceded phonological disambiguation. Subsequent analyses suggest that the Target preference is not the result of prosodic or co-articulatory cues to the Target noun. Instead it appears to reflect the online calculation of the scalar implicature.

In sum, the results from Experiment 2 suggest that there is a temporal lag between semantic and pragmatic influences on reference resolution in this task. When the lexical semantics of the quantifier is sufficient to disambiguate reference, participants make use of that information within 400 ms. When a scalar implicature is required, they fail to do so for at least 400 ms more. This is clearest in the contrast between *two* and *some*. In both cases, the referent must be disambiguated by recognizing that the Distractor is not compatible with an upper bound. In the case of *two*, we have argued that this upper bound is lexically encoded. Our findings suggest that it is available as rapidly as the lower boundaries for *three* and *all* (which are features of lexical semantics by all theoretical accounts). In the case of *some*, this upper bound is typically argued to result from a pragmatic implicature. Our data support this distinction and suggest that this implicature is considerably slower in its application.

There is however one alternate interpretation of this data which we have not addressed. Unlike the other terms, *some* is ambiguous between a lower-bounded and an upper-bounded reading. We have argued that this ambiguity reflects the presence or the absence of a scalar implicature. But perhaps the ambiguity is actually lexical in nature. For example, the two readings of *some* could be polysemous or homophonous words. If this were the case, then we might attribute the observed delay to difficulties in accessing the meaning of *some* rather than to sluggish use of scalar implicature. In typical cases of lexical ambiguity, both meanings of a word are initially activated but after a short delay only the contextually appropriate one persists (Swinney, 1979). If we assume that *some* has two meanings both of which are active, could this account for the observed delay in reference resolution? The answer depends on consequences of accessing two meanings.

One possibility is that the two competing interpretations of *some* both influence online semantic interpretation of the relative clause, restricting the reference of the noun that it modifies. The lower-bounded reading would be compatible with both the Target and the Distractor. The upper-bounded reading would be compatible with only the Target. Thus if both readings were equally preferred, we would still expect participants to look more at the Target. If the relative influence of each meaning was proportional to its frequency or contextual fit, then the Target preference would be even more robust (see Section 2.1 of Experiment 1 for evidence that the upper-bounded reading is preferred in this context and Section 5 for evidence that it is generally more common). However, in both Experiments 1 and 2 we found no evidence of a Target preference for *some* during the first 800 ms of the ambiguous region. Thus our data is inconsistent with this version of the ambiguity hypothesis.

An alternate possibility is that *some* has two meanings but the competition between them results in a stalemate that prevents participants from interpreting the relative clause and using it to restrict reference. This would result in precisely the pattern that we observed. Subjects would have no preference for either the Target or Distractor until the arrival of disambiguating phonological information. Experiment 3 tests this version of the ambiguity hypothesis. We reasoned that if lexical ambiguity on the *some* trials prevents participants from analyzing the relative clause, then there should be delays even when the Distractor is inconsistent with both the lower-bounded and upper-bounded interpretations (e.g., a girl with no socks or soccer balls). In contrast if the delay in the prior experiments reflected sluggish pragmatic processing, it should disappear when the semantics of the term is sufficient for reference resolution.

As in the previous experiments, participants in one set of *some* trials were presented with a girl that had some-but-not-all of the socks and another that had all of the soccer balls. We will be calling these "2-referent trials" because there are two referents that are consistent with the semantics of *some*. These trials were compared to a second set of *some* trials where participants were presented with a girl that had some-but-not-all of the socks and another that had none of the socks. We will be calling

Table 5
In Experiment 2, the probability of switching looks off the Target in Target initial trials for each quantifier type during each 100 ms time window

	Time from onset of the quantifier (in ms)								
	200	300	400	500	600	700	800	900	1000
Two	0.22 ($CI_{\pm 95}$ = .07)	0.23 ($CI_{\pm 95}$ = .07)	0.20 ($CI_{\pm 95}$ = .06)	0.11 ($CI_{\pm 95}$ = .04)	0.19 ($CI_{\pm 95}$ = .05)	0.18 ($CI_{\pm 95}$ = .06)	0.16 ($CI_{\pm 95}$ = .07)	0.17 ($CI_{\pm 95}$ = .07)	0.14 ($CI_{\pm 95}$ = .07)
Some	0.52 ($CI_{\pm 95}$ = .08)	0.55 ($CI_{\pm 95}$ = .08)	0.50 ($CI_{\pm 95}$ = .08)	0.41 ($CI_{\pm 95}$ = .09)	0.27 ($CI_{\pm 95}$ = .07)	0.21 ($CI_{\pm 95}$ = .06)	0.17 ($CI_{\pm 95}$ = .06)	0.08 ($CI_{\pm 95}$ = .03)	0.04 ($CI_{\pm 95}$ = .01)
Three	0.16 ($CI_{\pm 95}$ = .05)	0.13 ($CI_{\pm 95}$ = .04)	0.11 ($CI_{\pm 95}$ = .03)	0.09 ($CI_{\pm 95}$ = .03)	0.14 ($CI_{\pm 95}$ = .04)	0.11 ($CI_{\pm 95}$ = .05)	0.12 ($CI_{\pm 95}$ = .06)	0.12 ($CI_{\pm 95}$ = .05)	0.09 ($CI_{\pm 95}$ = .02)
All	0.18 ($CI_{\pm 95}$ = .05)	0.28 ($CI_{\pm 95}$ = .06)	0.28 ($CI_{\pm 95}$ = .06)	0.20 ($CI_{\pm 95}$ = .04)	0.09 ($CI_{\pm 95}$ = .03)	0.16 ($CI_{\pm 95}$ = .05)	0.23 ($CI_{\pm 95}$ = .07)	0.17 ($CI_{\pm 95}$ = .05)	0.12 ($CI_{\pm 95}$ = .05)
ANOVA (strength by scale)	$F1(1,16) = 4.59$, $p < .05$; $F2(1,15) = 7.73$, $p < .05$	$F1(1,16) = 1.99$, $p > .15$; $F2(1,15) = 2.89$, $p > .10$	$F1(1,16) = 1.22$, $p > .20$; $F2(1,15) = 2.03$ $p > .15$	$F1(1,16) = 2.98$, $p > .10$; $F2(1,15) = 5.29$, $p < .05$	$F1(1,16) = 2.37$, $p > .10$; $F2(1,15) = 0.74$, $p > .40$	$F1(1,16) = 0.02$, $p > .80$; $F2(1,15) = 0.11$, $p > .70$	$F1(1,16) = 0.56$, $p > .40$; $F2(1,15) = 0.02$, $p > .80$	$F1(1,16) = 1.79$, $p > .15$; $F2(1,15) = 0.24$, $p > .60$	$F1(1,16) = 2.17$, $p > .15$; $F2(1,15) = 0.37$, $p > .50$

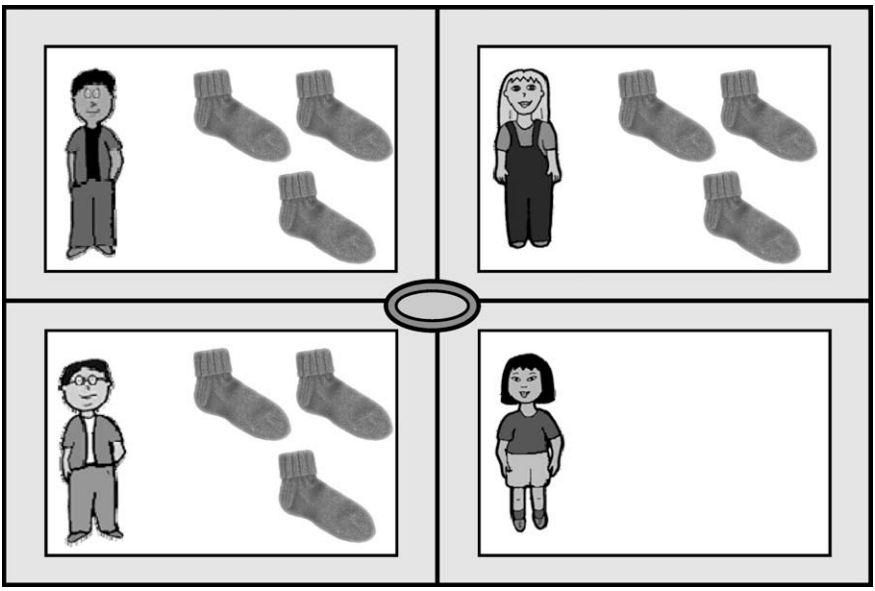

Fig. 8. Example of a visual-world display for *some*/*none* trials in Experiment 3. Participants here were instructed to "Point to the girl that has some of the socks."

these "1-referent trials" because there is only one referent that is consistent with the semantics of *some* (see Fig. 8). In the 1-referent trials, unlike the 2-referent trials, the Target can be resolved by the lower-bounded semantics of the quantifier alone, rather than by a pragmatic inference. Thus if pragmatic processing is delayed relative to semantic processing we should expect that participants will be considerably faster at disambiguating the Target of *some* in the 1-referent trials. If, however, lexical ambiguity accounts for the slower resolution of the referent in the *some* trials, and pragmatic inference is *not* involved, then this delay should be present in the 1-referent trials and we should expect similar performance for the two trial types.

4. Experiment 3

4.1. Methods

4.1.1. Participants

Twenty undergraduate students at Harvard University participated in this study. They received either course credit or $5 for their participation. All were native monolingual English speakers who had not participated in the previous experiments.

4.1.2. Procedure

The procedure was identical to the previous experiments.

4.1.3. Materials

The materials compared the interpretation of the scalar quantifier *some* in two different referential contexts. For the 2-referent trials, we again introduced participants to displays that contrasted a subset quantity of one item with the total set of another. Like previous experiments, participants saw two sets of objects distributed between boy–girl pairs using four stories like (8) above. These displays were presented with one minor modification: we equated the number of objects given to the Target and Distractor character in these critical trials in order to balance their visual salience. One set of six items

was split evenly between horizontally adjacent pair (girl with three socks and boy with three socks) and another set of three items which was given to one child from the remaining pair (girl with three soccer balls and boy with no soccer balls).

For the 1-referent trials, we introduced participants to displays that contrasted a subset quantity of one item with its empty set (see Fig. 8). Participants heard four new stories where a single set of objects were introduced and distributed among the boy–girl pairs like (10) below.

(10) The boys and girls on the soccer team were getting socks from the coach. The coach gave socks to Judy and socks to Craig and socks to Pat (*experimenter places three socks next to the girl on the upper-right, three socks next to the boy on the upper-left, and three socks next to the boy on the lower-left*). But these socks were too big for Cheryl's feet (*experimenter places a blank card next to the girl on the lower-right*).

Thus on these trials, three characters evenly shared nine items (girl and two boys with three socks) with a fourth character receiving nothing (girl with no socks).

We also included an equal number of filler trials to prevent participants from predicting the Target prior to the onset of the quantifier (see Appendix C for a list of all items). These filler trials used the same stories and displays as the 2-referent and 1-referent *some* trials above but used quantifiers that were consistent with the Distractor set. For the 2-referent displays, they were instead asked to select "the girl that has all of the socks" and for the 1-referent displays, they were asked for "the girl that has none of the socks." Like previous experiments, four items of each type were presented over the course of 16 randomized trials. The presentation of materials was counterbalanced by creating four lists such that each item appeared just once in every list and every item appeared in all four conditions across lists.

4.1.4. Coding

Data from the 2-referent and 1-referent trials were coded in the same manner as described in the previous experiments. Approximately 0.3% trials were excluded from further eye movement analyses because of participants' incorrect action responses. Approximately 0.3% trials were also excluded due to experimenter error. Finally, missing frames due to blinks or looks away accounted for 4% of all coded frames and were also excluded from analysis. There was 94.1% inter-coder reliability.

4.2. Results

4.2.1. Analyses of proportion of looking time to the Target

We examined the proportion of subjects' fixations towards the Target and Distractor characters over the same time windows specified in Experiments 1 and 2 (see Table 6 for the duration for each of these regions). Fixations to the other characters after onset of the gender cue were rare and accounted for less than 3% of total looks in the final three phases.

The period of most interest again was the Quantifier phase (*some of the soc-*) since a comparison between the 1- and 2-referent trials during this time gives us a direct measure of any delay in reference restriction via pragmatic inference. Based on our previous experiments, we would expect that when the subset is contrasted with the total set (2-referent trials), initial looks to the Target and Distractor would remain unbiased. On the other hand, we would predict that when the subset is contrasted with the empty set (1-referent trials), the semantic analysis of *some* would quickly rule out the Distractor and lead to a Target preference immediately following the onset of the quantifier.

Fig. 9 shows the proportion of looks to the Target for all four conditions. In some respects the looking patterns conform to our predictions. Prior to the onset of the quantifier, the Target preference was initially slightly below chance for both the 1-referent and 2-referent *some* trials, resulting in no significant differences between the two trial types during the Baseline and Gender phases (all t's < 1.50, all

Table 6
In Experiment 3, approximate duration of the time windows for the course-grained analysis

	Baseline phase	Gender phase	Quantifier phase	Disambiguation phase	End phase
Region length	533 ms	600 ms	733 ms	400 ms	733 ms

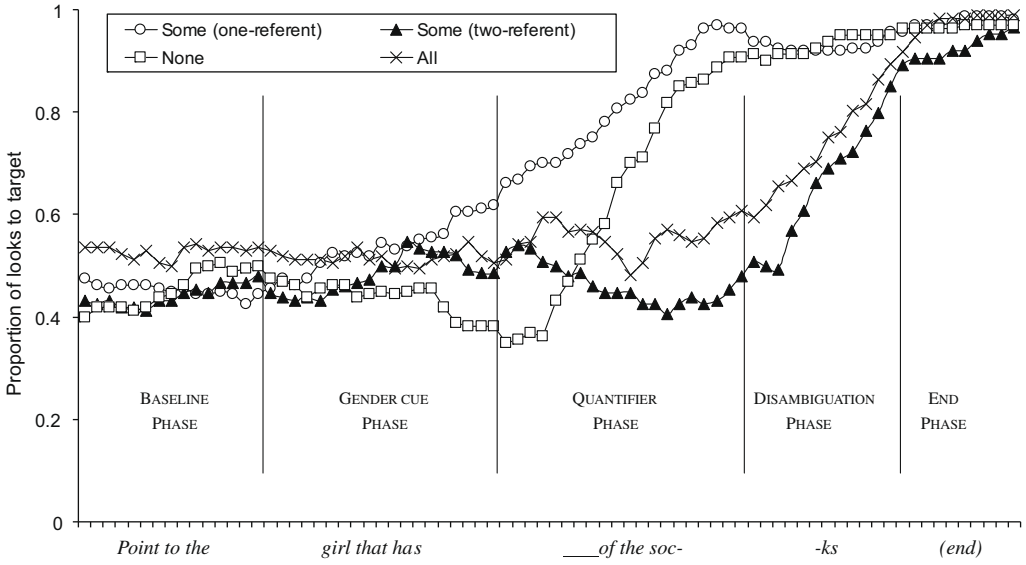

Fig. 9. In Experiment 3, the time-course of looks to Target for the four trial types.

p's > .15). This changed during the Quantifier phase: when *some* was contrasted with *none* fixations to the Target character increased (87%) but when *some* was contrasted with *all* looks to Target remained at chance (45%), resulting in a reliable difference between these conditions, $t1(19) = 6.48$, $p < .001$; $t2(15) = 7.58$, $p < .001$. This difference between 1-referent and 2-referent trials persisted throughout the Disambiguation phase (Target preference 95% and 74%, respectively) ($t1(19) = 5.01$, $p < .001$; $t2(15) = 3.77$, $p < .01$) and into the End phase where the difference in looks was significant by subjects ($t1(19) = 2.36$, $p < .05$) but not by items ($t2(15) = 1.41$, $p > .10$). However, since the Target fixations in this final period were near ceiling (99% in the 1-referent trials and 96% in the 2-referent trials), there was limited variability. Thus we followed up on the *t*-test with a non-parametric Wilcoxon signed-rank test and confirmed that the difference in Target preference during the End phase was significant by subjects ($W = 45$, $Z = 2.61$, $p < .005$) and marginally significant by items ($W = 19$, $Z = 1.68$, $p < .10$).

However, two features of this data complicate the interpretation of these analyses. First, there were systematic preferences for particular quantities prior to the quantifier onset. In the 1-referent display, looks to the Target in the *some* trials (57%) were significantly higher than in the *none* trials (40%) during the Gender phase, $t1(19) = 2.16$, $p < .05$; $t2(15) = 2.16$, $p < .05$. In contrast, looks in the 2-referent *some* trials (48%) were no different from looks in the *all* trials (51%) during the same period, $t1(19) = 1.48$, $p > .15$; $t2(15) = 1.05$, $p > .30$. This pattern suggests that prior to the onset of the quantifier participants preferred to look at characters with items rather than those with nothing.

Second, we found that looks to the Target in the *all* trials were slow to rise after the onset of the quantifier. Target preference during the Quantifier phase (55%) was significantly lower than in the comparable trials of Experiments 1 (66%) and 2 (72%), $F1(2,57) = 3.68$, $p < .05$; $F2(2,45) = 3.27$, $p < .05$. This could reflect differences in how the sets are construed across the two types of displays. In the *none* trials, participants heard instructions that quantify over a single set distributed among the four characters while in the *all* trials, there were two sets (socks and soccer balls) and the critical term only quantified over one of these sets. Perhaps, on some proportion of the *all* trials, participants were confused by this and attempted to interpret *all* as referring to the total set of objects. Since neither the Target nor the Distractor character had "all of the things" this should result in no reliable preference for either character. When these trials were averaged with trials in which participants limited the possible domains of quantification to the two basic level sets, we would expect to see a slight preference for the Target which would emerge more slowly across the trial.

4.2.2. Analyses of eye movements initiated after the quantifier

We again attempted to clarify how eye movements changed in response to the quantifier by dividing the trials into Non-Target initial and Target initial. In Experiment 2, we used this method of analysis to factor out initial perceptual biases. However, it has the added advantage of splitting the data in a way that allows us to isolate and examine two processes which may be distinct: the process of recognizing that the currently fixated object is the referent and holding fixation and the process of recognizing that the currently fixated object is not the correct referent and switching to the Target. For each of these subsets of data we analyzed looking time in 100 ms intervals from the onset of the quantifier until 1000 ms after quantifier onset (Fig. 10). Each time window was analyzed with one-way ANOVAs with all four trial types as within subject and item variables, and list/item group as a between subjects and items variable.

Table 7 illustrates that on the Non-Target initial trials, the pattern of fixations across the four trial types began to differ about 400 ms after the onset of the quantifiers ($F1(3,48) = 5.88$, $p < .01$; $F2(3,45) = 8.60$, $p < .001$). This effect is driven by switches to the Target in the *some* and *none* 1-referent conditions, suggesting that participants were able to rapidly use the semantics of the quantifier to rule out the Distractor character. In all of the time windows, there were no differences between the *some* and

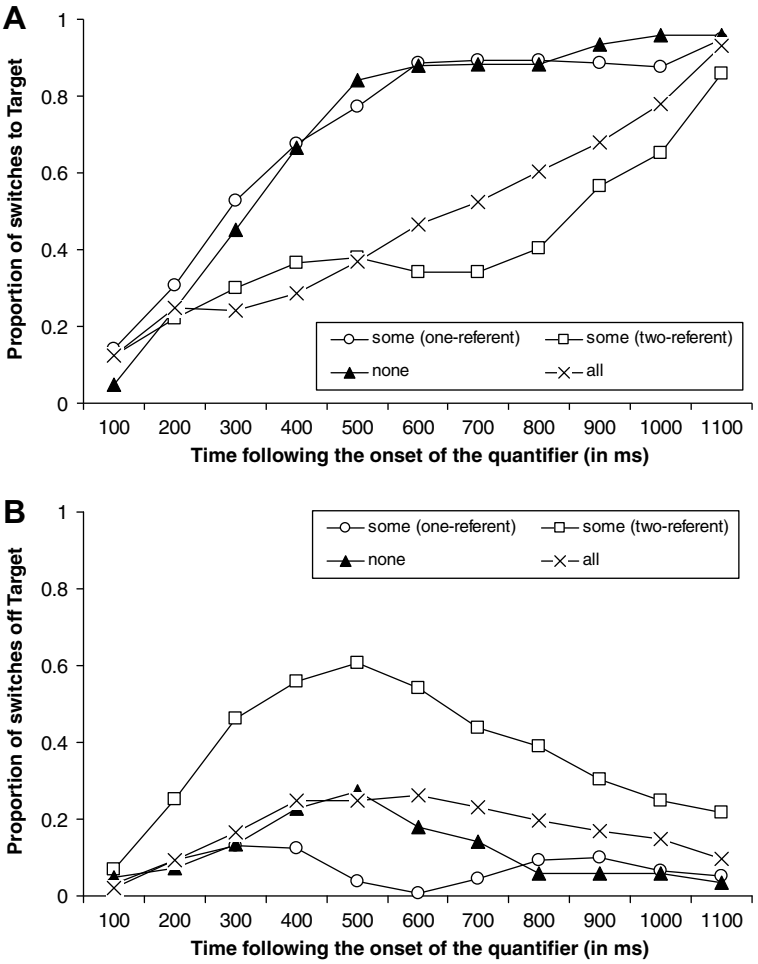

Fig. 10. Experiment 3, trials were separated based on fixations prior to the onset of the quantifier (A) Non-Target initial trials: the proportion of switches to the Target and (B) Target initial trials: the proportion of switches off the Target.

none 1-referent displays (all *p*'s > .15), demonstrating that there is no delay in interpreting *some* when meaning alone is sufficient for disambiguation. The pattern for the 2-referent displays was quite different. Participants were initially more likely to continue looking at a Non-Target item (typically the Distractor). In the early time windows, there were no reliable differences between *some* and *all*. This is surprising given the findings of Experiments 1 and 2 but consistent with our suggestion that participants may be uncertain about the domain of quantification in the 2-referent trials. Whatever the source of this confusion, by the 800 ms time window, participants in the *all* trials were marginally more likely to switch to the Target than those in the 2-referent *some* trial (*p*'s < .10), suggesting that the robust differences that we observed between *some* and *all* in our first experiments shape interpretation here as well. Critically, from the 400 ms to the 1000 ms time window, Target switches were more frequent for *some* in the 1-referent context than for *some* in the 2-referent context (all *p*'s < .05), demonstrating that lexical ambiguity alone cannot account for the delays observed for *some* in Experiments 1 and 2.

The Target initial trials provide stronger evidence that there is delay in calculating the upper bound of *some* (Table 8). The pattern of fixations across the four trial types began to differ about 300 ms after the onset of the quantifiers ($F1(3,48) = 3.70$, $p < .05$; $F2(3,45) = 7.14$, $p < .01$). This effect was driven by the fact that participants who heard *all, none*, and *some* in the 1-referent context rarely abandoned the Target, while those who heard *some* in the 2-referent context frequently did so. In contrast with the Non-Target analyses, we found a robust difference between *all* and 2-referent *some*, which emerged during the 300 ms time window and continued into the 600 ms time window. As in Experiment 2, participants were more likely to switch their gaze from the Target following *some* than *all* (all *p*'s < .05). Furthermore, from the 300 ms to the 800 ms time window, they were also more likely to switch following *some* in the 2-referent context than in the 1-referent context (all *p*'s < .05). In other words, when participants could rule out the Distractor using semantics, they typically continued to fixate the Target. But when the Distractor was compatible with the semantics of the quantifier they often shifted their gaze. Critically, this study demonstrates a gap between semantic and pragmatic analysis of the same term: the meaning of *some* restricts reference shortly after quantifier onset, while the implicature generated by *some* does not.

4.3. Discussion

In Experiment 3, we again replicated delays in looks to the Target for critical trials that contrasted *some* with a total set (2-referent trials). However, similar delays were not seen in trials that contrasted *some* with an empty set (1-referent trials). This pattern was confirmed when we separated trials by initial fixations and found that in the 1-referent case, participants were more likely to switch their looks to the correct Target and less likely to switch to an incorrect Non-Target following the onset of *some*. These results suggest that resolution of the Target is quicker via semantic analysis than pragmatic inference.

One might wonder whether the early emerging Target preference in the 1-referent *some* simply reflects a general lack of interest in looking at a character with no objects (a preference for looking at a girl with three socks rather than one with no socks). Such a preference would have no effect on the 2-referent trials since Targets and Distractors in those displays were matched for number of objects (three socks vs. three soccer balls). If this were the case, then we would expect that participants would favor characters with objects over those without objects regardless of what was said. However, in the *none* trials, participants rapidly converged on the Target character (the girl with none of the socks) resulting in a reliable preference for the Target (69%, $t1(19) = 6.17$, $p < .001$; $t2(15) = 5.54$, $p < .001$) during the critical Quantifier phase.

The *all* trials in Experiment 3 produced a data pattern that was somewhat different than what we saw in Experiments 1 and 2. Since the displays and utterances were virtually identical, we tentatively suggest that the inclusion of the 1-referent scenes opened up the possibility that the domain of quantification would be the entire set of objects. By splitting the trials into Target initial and Non-Target initial trials, we determined that the sluggish rise in Target preference for *all* was largely driven by a failure to switch to Target on Non-Target initial trials. This provides a tantalizing suggestion about how and when a broader domain of quantification will be considered. On Non-Target initial trials, participants have to consider both the Distractor set (to reject it) and

the Target set (to shift), making them perhaps more likely to consider the total set of objects as a potential domain of quantification.

While the performance on the *all* trials was slower in this experiment than in the previous ones, we still see evidence for an asymmetry between understanding that *all* must refer to the total set and inferring that phrases with *some* should not. In Non-Target initial trials this is apparent in the 800 ms time window where see more switches to the Target on *all* trials relative to 2-referent *some*. In the Target initial trials, this asymmetry emerges much earlier (300–700 ms): participants hearing *some* often switch their gaze to the Distractor character with the total set but participants hearing *all* rarely switch gaze to the character with the subset.

Critically, Experiment 3 demonstrates that *some* is interpreted quite quickly in the 1-referent context where the semantics of the term are sufficient for reference resolution. Thus the delays that we saw for *some* in Experiments 1 and 2 cannot be attributed to processing stalemate brought on by lexical ambiguity. All together, this pattern is consistent with an account of language processing where semantic analysis begins before pragmatic interpretation. Initial analysis of the lower-bounded semantics of *some* leads to competition between the Target and Distractor when it contrasts with *all* but not when it contrasts with *none*.

5. General discussion

This study explores the real-time interaction between semantic and pragmatic meaning by investigating the interpretation of scalar terms. In Experiment 1 and 2, we found quick resolution of the referent when participants heard *two, three*, and *all* but initial delays when they heard *some* and had to make an upper-bounding inference. In Experiment 3, we again found delays in looks to the referent when *some* contrasted with *all* but we found no such delays when *some* was contrasted with *none*. These findings add to a growing literature demonstrating that implicature requires additional processing which begins only after semantic analysis is underway (Bott & Noveck, 2004; Breheny et al., 2006; De Neys & Schaeken, 2007; Noveck & Posada, 2003). This study sheds light on the nature of this delay by elucidating how the quantifier is interpreted as the analysis progresses. The results are consistent with a model of language processing where the information provided by distinct levels of representation becomes available at different times over the course of real-time comprehension.

In the remainder of this discussion, we will examine three issues. First, we will introduce and address some alternate interpretations of these findings. Next, we will explore how our findings bear on theories of the relation between semantic and pragmatic processing in general as well as the interpretation of scalar quantifiers in particular. Finally, we will address an apparent tension between our results and previous findings in psycholinguistics demonstrating rapid assimilation of extra-linguistic cues (e.g., Sedivy, 2003; Sedivy et al., 1999).

5.1. Alternate hypotheses for the delay in interpreting "*some*"

Could the delays that were observed in the *some* conditions be accounted for by another process besides scalar implicature? In this section, we address two alternate possibilities.

First, one might argue that the differences between the gaze time patterns for *some* and *two* in Experiments 1 and 2 are attributable to differences in the verification conditions for numbers and scalar quantifiers. The applicability of a number can be verified solely by looking at the set of objects owned by the character in question (*Does the girl have exactly two socks*?). In contrast, to determine whether the upper-bounded reading of *some* applies, a participant would have to examine both the set of objects belonging to the character in question and the set of those objects belonging to the adjacent character (*Does the girl have some socks and does the adjacent boy have at least one as well*?). This might require additional processing and perhaps additional eye movements as well, explaining why looks to the Target were slower for *some* than for *two*.

We think this account is unlikely for three reasons. First, it predicts, incorrectly, that we should see parallel delays for *all*. Like *some*, the verification conditions for *all* involve the set of objects that the Target has and the set owned by the adjacent character (*Does the girl have some socks and the does the boy have*

Table 7
In Experiment 3, the probability of switching looks to the Target in the Non-Target initial trials for each quantifier type during each 100 ms time window

	Time from onset of the quantifier (in ms)								
	200	300	400	500	600	700	800	900	1000
Some (1-referent)	0.30 ($CI_{\pm 95}$ = .07)	0.52 ($CI_{\pm 95}$ = .06)	0.67 ($CI_{\pm 95}$ = .07)	0.76 ($CI_{\pm 95}$ = .07)	0.87 ($CI_{\pm 95}$ = .06)	0.88 ($CI_{\pm 95}$ = .06)	0.88 ($CI_{\pm 95}$ = .06)	0.88 ($CI_{\pm 95}$ = .06)	0.87 ($CI_{\pm 95}$ = .05)
Some (2-referent)	0.22 ($CI_{\pm 95}$ = .07)	0.30 ($CI_{\pm 95}$ = .09)	0.36 ($CI_{\pm 95}$ = .09)	0.38 ($CI_{\pm 95}$ = .09)	0.34 ($CI_{\pm 95}$ = .08)	0.34 ($CI_{\pm 95}$ = .09)	0.40 ($CI_{\pm 95}$ = .07)	0.56 ($CI_{\pm 95}$ = .07)	0.65 ($CI_{\pm 95}$ = .08)
None	0.24 ($CI_{\pm 95}$ = .05)	0.45 ($CI_{\pm 95}$ = .07)	0.67 ($CI_{\pm 95}$ = .06)	0.84 ($CI_{\pm 95}$ = .05)	0.88 ($CI_{\pm 95}$ = .04)	0.88 ($CI_{\pm 95}$ = .04)	0.88 ($CI_{\pm 95}$ = .06)	0.93 ($CI_{\pm 95}$ = .04)	0.96 ($CI_{\pm 95}$ = .03)
All	0.25 ($CI_{\pm 95}$ = .08)	0.25 ($CI_{\pm 95}$ = .08)	0.30 ($CI_{\pm 95}$ = .08)	0.38 ($CI_{\pm 95}$ = .08)	0.47 ($CI_{\pm 95}$ = .07)	0.53 ($CI_{\pm 95}$ = .08)	0.61 ($CI_{\pm 95}$ = .08)	0.68 ($CI_{\pm 95}$ = .07)	0.78 ($CI_{\pm 95}$ = .07)
ANOVA	$F1(3,48) = 0.21$, $p > .80$; $F2(3,45) = 0.98$, $p > .40$	$F1(3,48) = 2.63$, $p < .10$; $F2(3,45) = 3.86$, $p < .05$	$F1(3,48) = 5.88$, $p < .05$; $F2(3,45) = 8.60$, $p < .01$	$F1(3,48) = 8.95$, $p < .01$; $F2(3,45) = 15.81$, $p < .01$	$F1(3,48) = 15.21$, $p < .01$; $F2(3,45) = 22.86$, $p < .01$	$F1(3,48) = 14.87$, $p < .01$; $F2(3,45) = 16.71$, $p < .01$	$F1(3,48) = 13.13$, $p < .01$; $F2(3,45) = 11.75$, $p < .01$	$F1(3,48) = 10.16$, $p < .01$; $F2(3,45) = 7.09$, $p < .01$	$F1(3,48) = 5.95$, $p < .01$; $F2(3,45) = 5.44$, $p < .01$
t-Test (1- vs. 2-referent Some)	$t1(19) = 0.67$, $p > .50$; $t2(15) = 1.49$, $p > .15$	$t1(19) = 1.79$, $p < .10$; $t2(15) = 2.41$, $p < .05$	$t1(19) = 2.38$, $p < .05$; $t2(15) = 3.47$, $p < .01$	$t1(19) = 2.86$, $p < .05$; $t2(15) = 5.46$, $p < .001$	$t1(19) = 4.14$, $p < .01$; $t2(15) = 9.14$, $p < .001$	$t1(19) = 4.64$, $p < .001$; $t2(15) = 8.48$, $p < .001$	$t1(19) = 4.76$, $p < .001$; $t2(15) = 5.04$, $p < .001$	$t1(19) = 3.52$, $p < .01$; $t2(15) = 2.68$, $p < .05$	$t1(19) = 2.75$, $p < .05$; $t2(15) = 2.30$, $p < .05$
t-Test (All vs. 2-referent Some)	$t1(19) = 0.22$, $p > .80$; $t2(15) = 0.47$, $p > .60$	$t1(19) = 0.44$, $p > .60$; $t2(15) = 0.83$, $p > .40$	$t1(19) = 0.53$, $p > .60$; $t2(15) = 0.38$, $p > .70$	$t1(19) = 0.02$, $p > .90$; $t2(15) = 0.58$, $p > .50$	$t1(19) = 1.14$, $p > .20$; $t2(15) = 1.89$, $p < .10$	$t1(19) = 1.59$, $p > .10$; $t2(15) = 1.65$, $p > .10$	$t1(19) = 1.97$, $p < .10$; $t2(15) = 1.83$, $p < .10$	$t1(19) = 1.14$, $p > .20$; $t2(15) = 1.43$, $p > .15$	$t1(19) = 1.16$, $p > .20$; $t2(15) = 1.69$, $p > .10$

Table 8
In Experiment 3, the probability of switching looks off the Target in Target initial trials for each quantifier type during each 100 ms time window

	Time from onset of the quantifier (in ms)								
	200	300	400	500	600	700	800	900	1000
Some (1-referent)	0.09 ($CI_{\pm 95}$ = .05)	0.13 ($CI_{\pm 95}$ = .07)	0.13 ($CI_{\pm 95}$ = .06)	0.04 ($CI_{\pm 95}$ = .03)	0.01 ($CI_{\pm 95}$ = .01)	0.04 ($CI_{\pm 95}$ = .03)	0.09 ($CI_{\pm 95}$ = .04)	0.10 ($CI_{\pm 95}$ = .05)	0.07 ($CI_{\pm 95}$ = .04)
Some (2-referent)	0.25 ($CI_{\pm 95}$ = .06)	0.46 ($CI_{\pm 95}$ = .09)	0.56 ($CI_{\pm 95}$ = .09)	0.61 ($CI_{\pm 95}$ = .09)	0.54 ($CI_{\pm 95}$ = .08)	0.44 ($CI_{\pm 95}$ = .08)	0.39 ($CI_{\pm 95}$ = .08)	0.30 ($CI_{\pm 95}$ = .08)	0.25 ($CI_{\pm 95}$ = .08)
None	0.07 ($CI_{\pm 95}$ = .03)	0.13 ($CI_{\pm 95}$ = .04)	0.23 ($CI_{\pm 95}$ = .06)	0.27 ($CI_{\pm 95}$ = .08)	0.18 ($CI_{\pm 95}$ = .06)	0.14 ($CI_{\pm 95}$ = .05)	0.06 ($CI_{\pm 95}$ = .03)	0.06 ($CI_{\pm 95}$ = .03)	0.06 ($CI_{\pm 95}$ = .03)
All	0.09 ($CI_{\pm 95}$ = .05)	0.17 ($CI_{\pm 95}$ = .07)	0.25 ($CI_{\pm 95}$ = .08)	0.25 ($CI_{\pm 95}$ = .08)	0.26 ($CI_{\pm 95}$ = .08)	0.23 ($CI_{\pm 95}$ = .06)	0.20 ($CI_{\pm 95}$ = .06)	0.17 ($CI_{\pm 95}$ = .06)	0.15 ($CI_{\pm 95}$ = .06)
ANOVA	$F1(3,48) = 2.01$, $p > .10$; $F2(3,45) = 3.00$, $p < .05$	$F1(3,48) = 3.70$, $p < .05$; $F2(3,45) = 7.14$, $p < .01$	$F1(3,48) = 4.50$, p<.01; $F2(3,45) = 6.56$, $p < .01$	$F1(3,48) = 8.10$, $p < .01$; $F2(3,45) = 7.82$, $p < .01$	$F1(3,48) = 10.07$, $p < .01$; $F2(3,45) = 7.15$, $p < .01$	$F1(3,48) = 5.65$, $p < .01$; $F2(3,45) = 4.70$, $p < .01$	$F1(3,48) = 4.41$, $p < .01$; $F2(3,45) = 5.38$, $p < .01$	$F1(3,48) = 2.09$, $p > .10$; $F2(3,45) = 3.74$, $p < .05$	$F1(3,48) = 1.64$, $p > .15$; $F2(3,45) = 2.84$, $p < .05$
t-Test (1- vs. 2-referent Some)	$t1(19) = 1.63$, $p > .10$; $t2(15) = 2.22$, $p < .05$	$t1(19) = 2.64$, $p < .05$; $t2(15) = 3.43$, $p < .01$	$t1(19) = 3.33$, $p < .01$; $t2(15) = 3.56$, $p < .01$	$t1(19) = 5.28$, $p < .001$; $t2(15) = 4.53$, $p < .001$	$t1(19) = 6.18$, $p < .001$; $t2(15) = 4.88$, $p < .001$	$t1(19) = 4.21$, $p < .001$; $t2(15) = 3.63$, $p < .01$	$t1(19) = 2.88$, $p < .01$; $t2(15) = 2.79$, $p < .05$	$t1(19) = 1.96$, $p < .10$; $t2(15) = 2.14$, $p < .05$	$t1(19) = 1.78$, $p < .10$; $t2(15) = 1.91$, $p < .10$
t-Test (All vs. 2-referent Some)	$t1(19) = 1.92$, $p < .10$; $t2(15) = 2.29$, $p < .05$	$t1(19) = 2.59$, $p < .05$; $t2(15) = 3.02$, $p < .01$	$t1(19) = 2.37$, $p < .05$; $t2(15) = 2.66$, $p < .05$	$t1(19) = 2.76$, $p < .05$; $t2(15) = 2.89$, $p < .05$	$t1(19) = 2.21$, $p < .05$; $t2(15) = 1.96$, $p < .10$	$t1(19) = 1.84$, $p < .10$; $t2(15) = 2.55$, $p < .05$	$t1(19) = 1.78$, $p < .10$; $t2(15) = 2.42$, $p < .05$	$t1(19) = 1.07$, $p > .20$; $t2(15) = 2.11$, $p < .10$	$t1(19) = 0.78$, $p > .40$; $t2(15) = 1.72$, $p > .10$

none?). Yet in both Experiments 1 and 2, the referent for *all* was disambiguated as rapidly as the referent of *three*, suggesting that this difference created no measurable delay. Second, the fact that we observed a delay for *three* in Experiment 1, in which *three* described the total set, but not in Experiment 2, in which *three* described a subset of items, provides indirect evidence that participants were sensitive to the distribution of objects across characters even on trials in which a number was used. Finally, the participants' eye movements do not suggest that were overtly verifying the sentences in this way. Looks to the adjacent character were vanishingly rare in all conditions and accounted for less than 4% of total looks after the onset of the quantifier across all three experiments. Furthermore, the proportion of such fixations did not differ across conditions, suggesting no privileged strategy for the *some* trials relative to other quantifiers (all p's > .20).

A second alternate hypothesis returns to the observation that *some*, unlike the other terms, is ambiguous between a lower-bounded and an upper-bounded reading. As we noted earlier, this ambiguity could be attributed to two polysemous forms, rather than a single form plus a defeasible inference. Experiment 3 rules out one version of the ambiguity hypothesis: clearly the putative ambiguity does not stall interpretation leading to a delay in reference restriction whenever *some* is used. Instead the delay is limited to cases in which both referents are compatible with the semantics of *some*. However, there is another version of the ambiguity hypothesis which we alluded to earlier. Perhaps during word recognition, polysemous words compete in much the same way as words with phonological overlap or homophonous words. If this were the case then we would expect that during ambiguous period, participants would look at the referents that are consistent with either of the two meanings (Allopenna et al., 1998). This incremental version of the ambiguity hypothesis account makes two correct predictions. First, when *some* is pitted against *all* there should be fixations on both the Target and Distractor since both are compatible with the lower-bounded meaning. Second, when *some* is pitted against *none*, disambiguation should be rapid since the character with no objects is inconsistent with both readings.

However, this incremental ambiguity hypothesis fails to account for another salient feature of our data. If both meanings are accessed and influence reference resolution, we should still expect to see an early preference for the Target character on the critical trials in which *some* is pitted against *all*. The Target is compatible with both the lower-bounded reading and the upper-bounded reading, while the Distractor is only compatible with the lower-bounded reading. Thus if the two readings were equally weighted and there were no other constraints on reference resolution, we would predict that looking time would be split 75–25 in favor of the Target. A fully-fleshed out version of the incremental ambiguity hypothesis would likely predict an even stronger bias for Target fixations since lexical processing, as measured by eye movements to potential referents, is strongly and rapidly influenced by lexical frequency (Dahan, Magnuson, & Tanenhaus, 2001) and task constraints (Mirman, Magnuson, Strauss, & Dixon, 2008). In our task, both of these factors should favor the upper-bounded reading and hence looks to the Target: in ordinary conversation, *some* generally implies *not all* (see Hamilton, 1860; Levinson, 2000; Papafragou & Musolino, 2003, inter alia) and within these experiments, *some* always ultimately referred to a proper subset. But in all three experiments, there was no sign of a preference for the Target until at least 800 ms after quantifier onset. This feature of the data is inconsistent with the incremental ambiguity hypothesis but completely consistent with our hypothesis that initial fixations reflect a single lower-bounded meaning which is later enriched to derive the upper-bounded interpretation.

5.2. Linguistic theories of scalar implicature

Within theoretical linguistics, there has been a long-standing controversy about the proper characterization of the relation between semantics and pragmatics. These border wars have centered on the questions of where semantics ends and pragmatics begins and how pragmatic inferences are calculated (Berg, 2002; Bezuidenhout & Cutting, 2002; Gibbs & Moise, 1997; Nicolle & Clark, 1999). Three types of theories are of particular interest for the psycholinguistic study of scalar implicature.

The first is the Neo-Gricean account. Traditional Gricean and Neo-Gricean theories state that literal sentence meanings are captured by well-formed semantic representations corresponding to logical forms (Horn, 1989, 1992; Levinson, 1983, 2000). In a classic Gricean account, scalar implicature results from a post-sentential process involving the application of communicative principles like the Quantity

Maxim (Grice, 1975, see Section 1). More recently, Neo-Gricean theorists have suggested that habitual use of implicatures could result in their automatization (Gadzar, 1979; Horn, 1984; Levinson, 1983, 2000). While this default interpretation can be canceled in the presence of conflicting evidence in context, cancellation occurs only after the scalar implicature has been generated. In a strong version of this hypothesis, the restricted meaning that results from the application of the implicature could be stored with that form in the lexicon.

A second proposal links scalar implicatures to grammatical properties of the sentence (Chierchia, 2004a). This account has several features in common with the Gricean proposal. In particular, there is a full semantic form which exist independent of and prior to the scalar inference and implicatures involve the enrichment or strengthening of this representation. Furthermore, these inferences are defeasible and the lower-bounded reading results from the absence or cancellation of this inference. However in Chierchia's theory, the generation of implicatures is prompted by the semantic structures in which the scalar terms appear. In most grammatical contexts, scalar implicatures are calculated *locally*—in the same clause as the quantifier. However, some semantic structures (downward-entailing contexts) create environments in which a scalar implicature would result in a weaker, less restrictive utterance (e.g., negation or an if-clause). Since this information is available within the clause, it can prevent the implicature from being calculated.

Finally, Relevance theory, in contrast with Gricean and grammatical theories, proposes that many features of semantics are inherently underspecified. This underspecification motivates a theory in which pragmatics is a constructive process that draws on global knowledge of the situation to flesh out these skeletal semantic representations (Carston, 1998; Recanati, 2003; Sperber & Wilson, 1986/1995). Relevance theorists reject the notion of default inferences and instead suggest that all pragmatic interpretation including scalar implicature is based on a more general principle of relevance. For any given linguistic message, listeners engage in inferential processing until they have met some internal criterion for the relevance of the message. This establishes a tradeoff between the possible gains associated with generating an inference and the amount of cognitive effort necessary to derive it. As a result, scalar implicatures will only be generated when they are required to meet the listener's internal standard of relevance.

While these accounts draw on distinctions that have psychological implications, they are theories of linguistic representation and not of language processing (see Bezuidenhout & Cutting, 2002). Consequently, there is no straightforward mapping between these theories and patterns of performance in real-time comprehension. For example, notions of automaticity in the Neo-Gricean sense do not specify whether the scalar implicature is generated immediately upon hearing the term or merely prior to conscious awareness. This concept typically conflates the notion of rapidity (How quickly are implicatures calculated?) with the notion of defaultness (Are they always generated?). Similarly, the concept of relevance, as used in Relevance theory, does not specify how or when various forms of non-linguistic information, such as the speaker's intention or discourse context, are actually integrated in sentence processing.

Nonetheless, we believe that our data speak to aspects of these theories. First and most directly, our data impose constraints on the automatic processes invoked in the Neo-Gricean theories and to a lesser extent, in the grammatical theory. Our data present a particular challenge to Levinson's proposal (2000): if scalar implicatures are lexicalized, then we might expect them to emerge as rapidly as semantically-encoded upper-bounds. Our experiments employ the paradigmatic scalar quantifier *some* within a structural context where the implicature should be licensed by the grammar. Yet we found a delay of about 400 ms between the use of the lexically encoded upper-bound of *two* and the pragmatically inferred upper-bound of *some*. Apparently, even the most robust pragmatic inferences take additional time to compute.[5]

[5] Our experiments, however, do not rule the possibility that scalar implicatures are calculated by default since the interpretation of *some* was only examined when it was consistent with the implicature. Future work will examine patterns of comprehension in a context where the implicature interpretation is actually violated, i.e. when *some* is used to refer to the total set. Here, we would predict that if implicatures were calculated by default during real-time comprehension, then violation of these inferences would cause delays in language processing.

The apparent presence of an online implicature may impose complementary constraints on the more open-ended processes invoked by Relevance theory and the Classical Gricean account. Recall that in Experiment 2, we found that participants in the *some* trials showed a reliable preference for the Target prior to phonological disambiguation. This suggests that after a period of semantic analysis, listeners arrived at an upper-bounded interpretation by generating a scalar implicature during the course of real-time processing. The speed with which this inference was calculated cannot rule out the possibility that listeners engage deliberative counterfactual reasoning, as the Classical Gricean theory would suggest. However, these results do suggest that any deliberations of this kind are concluded quickly, at least under these circumstances. For Relevance theory, our data presents something of a paradox. In the framework of Relevance theory, the calculation of a pragmatic inference depends on the tradeoff between cognitive effort and communicative gain. In our task there was little to be gained by making the inference: after all, the referent of the quantified phrase was always lexically disambiguated, making the implicature unnecessary for full comprehension. From a perspective of Relevance theory, the fact that the scalar inference was nevertheless calculated suggests that it must have a fairly low cost, perhaps because it is so frequently deployed. Thus, these data suggest that Relevance theory may need to acknowledge the possibility that scalar implicatures have a preferred status relative to other pragmatic inferences.

5.3. Online processing and the interface between semantics and pragmatics

In this experiment, we used scalar implicature to explore the temporal relation between semantic and pragmatic processing. Many prior studies have asked parallel questions by examining how referential information is used to resolve syntactic and referential ambiguity during language comprehension (Arnold, Eisenband, Brown-Schmidt, & Trueswell, 2000; Hanna, Tanenhaus, & Trueswell, 2003; Nadig & Sedivy, 2002; Tanenhaus et al., 1995). One prominent line of work that is relevant to our discussion is Sedivy's studies of adjective comprehension (Grodner & Sedivy, in press; Sedivy, 2003; Sedivy et al., 1999). Sedivy and colleagues (1999) found that when participants were instructed to "Pick up the tall glass," they identified the correct target faster and made fewer spurious looks to a competing item (*tall pitcher*) in the presence of a contrast object (*short glass*). Sedivy (2003) suggests that the presence of this contrast item leads to a rapid Gricean inference that restricts reference in the following way: when listeners hear *tall* they infer that it probably modifies a member of a contrastive set, since the adjectival modification would not be necessary if the item could be uniquely identified from the noun alone. Thus when a single contrastive set is present in the scene, they can use this information to identify the referent. The fact that this process appears to occur prior to the onset of the noun has led the authors to conclude that these results "present clear evidence for the interaction of contextual and linguistic information at the earliest possible moments" (Sedivy et al., 1999, pp. 143).

On this construal, there is a tension between the Sedivy results and our current findings: Why would one pragmatic inference be delayed, when another similar one is so quick? Here, we explore four alternate explanations for the apparent discrepancy.

The first possibility is that the two lines of work do in fact tap fundamentally different processes. For example, perhaps the semantics of scalar adjectives includes a contextual parameter that incorporates referential information, while the interpretation of scalar quantifiers involves a post-semantic pragmatic inference of the kind envisioned by Grice. An explanation of this kind would both remove the apparent discrepancy and raise new questions about the nature of these two kinds of pragmatic effects and their place in the linguistic system.

The second possibility is that the timing of semantic and pragmatic processes in the two studies is similar despite appearances to the contrary. Perhaps in the case of scalar adjectives pragmatic inference is also preceded by a short period of semantic analysis. This would be consistent with most linguistic theories which assert that the relevance of the contrast set for predicting the reference of the noun phrase can only be established after the system has accessed the semantic properties of adjective itself. For example in the case of *tall*, we must recognize that it is a modifier which stipulates height with respect to a comparison class before we can recognize the relevance of the contrast item. In the absence of this information, the presence of the short glass is uninfor-

mative (contrast with "Pick up the red glass" in the presence of short and tall glasses both of which are red). Because the Sedivy experiments do not compare the timing of context effects with the timing of semantic processing, the studies do not rule out the possibility that the contextual inference is preceded by rapid analysis of the semantic features of the adjective itself.

The third possibility is that delays that we observed in our study do not reflect general features of pragmatic interpretation, but are simply attributable to features of our contexts or utterances, which made it more difficult for participants in our study to make the upper-bounding inference. Perhaps participants may have failed to encode the how each set of objects was divided between the characters, making them unable to use this contrast in inferential processing. However, this option seems unlikely since listeners were able to effectively use the contrasting set to rule out semantically inconsistent Distractors (i.e. in the *all* trials and 1-referent *some* trials). In addition, when participants were given the same stories and displays, they readily generated the scalar implicature in an offline task.

The final possibility is that the scalar inference is less robust for scalar quantifiers than it is for the scalar adjectives. To explore this hypothesis, we examined a random sample of 50 usages of *some* and *tall* from the The British National Corpus (BNC). For *some*, we looked for cases that unambiguously referred to a subset and for *tall*, we looked for cases that unambiguously contrast the height of two (or more) referents. If the scalar inference is less robust for quantifiers compared to adjectives, then we would expect to find very few subset interpretations for *some* but many overt height comparisons for *tall*. Instead we found the opposite pattern. There were in fact many instances where *some* clearly referred to a subset like in (11).

(11) The trust plans to replace some of Moncrieffe Hill's conifers with broadleaf trees.

This interpretation accounted for 42% of the sentences, demonstrating that the upper-bounded inference is often associated with interpretation of this quantifier. In contrast, we failed to find any examples that appeared to refer to a total set, suggesting that the lower-bounded interpretation may be vanishingly rare in real-world communication. For *tall*, we found relatively few cases of overt comparisons between two potential referents. Sentences like (12) occurred only 10% of the time. Instead usages like (13), which referenced an implicit standard of height, were far more common. These interpretations accounted for 74% of the sentences.[6]

(12) Tall bamboos wave above dwarf species.
(13) He was a tall, hawk-like figure, noted for certain eccentricities.

Altogether, these patterns do not support the notion that scalar inferences are weaker for quantifiers compared to adjectives. If anything, we found tentative evidence that the implicature was more robust for *some* than for *tall*.

5.4. Conclusion

In three experiments, we found evidence of a temporal lag between semantic processing and the initiation of pragmatic processing. When the semantics of the quantifier disambiguated the referent, listeners quickly restricted reference to the correct target. However, when presented with the lexically lower-bounded quantifier *some*, they initially failed to generate an implicature that would rule out the total set. Thus our results suggest that while scalar inferences may be rapid, they are preceded by some degree of semantic analysis. This is consistent with a model of linguistic architecture where semantic representations serve as a mediator between phonological form and pragmatic interpretation.

[6] Many usages of *some* and *tall* fell outside our primary categories of interest. *Some* was used as an indefinite determiner in 40% of the sentences ("Can I have some water?") and adopted various other usages like superlative ("That was some party!") or estimation ("There were some 80 people in the room") in the remaining 18% of the cases. *Tall* appeared in a measurement phrase ("He was five-feet-tall") 16% of the time.

Acknowledgments

This work benefited from conversations with members of the Laboratory for Developmental Studies and MIT-Harvard Number Reading Group. We are grateful to Ayo Adigun, Hila Katz, Charlotte Distefano, and Jane Pollock for their assistance in data collection and coding. Portions of this work have been presented at the 19th annual meeting of CUNY Sentence Processing and the 28th annual meeting of the Cognitive Science Society. This material is based upon work supported by the National Science Foundation under Grant No. 0623845.

Appendix A. Stimuli items for Experiments 1 and 2

Item	Instruction	Distracter
1	Point to the girl that has *some/two/three/all* of the matches	Maps
2	Point to the boy that has *some/two/three/all* of radios	Rain clouds
3	Point to the girl that has *some/two/three/all* of the sandals	Sandwiches
4	Point to the boy that has *some/two/three/all* of the roofs	Roosters
5	Point to the girl that has *some/two/three/all* of the rats	Rabbits
6	Point to the girl that has *some/two/three/all* of the pills	Pillows
7	Point to the boy that has *some/two/three/all* of the cards	Cars
8	Point to the girl that has *some/two/three/all* of the watermelons	Waffles
9	Point to the boy that has *some/two/three/all* of the snails	Snakes
10	Point to the girl that has *some/two/three/all* of the dogs	Dolls
11	Point to the boy that has *some/two/three/all* of the mushroom	Muffins
12	Point to the boy that has *some/two/three/all* of the peas	Pizzas
13	Point to the boy that has *some/two/three/all* of the bees	Beetles
14	Point to the girl that has *some/two/three/all* of the candles	Candies
15	Point to the girl that has *some/two/three/all* of the socks	Soccer balls
16	Point to the boy that has *some/two/three/all* of the robes	Roses

Appendix B. Assessing the prosodic structure and felicity of the critical utterances

To ensure that the prosody of the target utterances was natural and consistent across the four item types, we conducted two analyses. First, to obtain a more detailed description of the prosodic structure of the recorded instructions, we had a trained research assistant code the critical utterances used in Experiments 1 and 2 using the ToBI annotation system (Beckman & Hirschberg, 1994). Second, we administered a rating task to verify the prosodic felicity of these utterances in the context in which they appeared. We had believed that utterances without contrastive or focal stress on either the gender cue or the quantifier would be most natural for all four of the quantifiers given these displays. To verify this, we asked naïve participants to rate the utterances from the original experiment and alternate prosodic versions of these utterances with focal stress on the either the gender or the quantifier.

B.1. ToBI analysis

A trained research assistant coded the prosodic structure of the original sentences using the ToBI annotation system (Beckman & Hirschberg, 1994). We were particularly interested in two relevant features of the utterances. First, we wanted to evaluate the type and consistency of the pitch accents on each of the content words (e.g., point, girl, some, socks). Second, we wanted to evaluate the presence of the break index between the quantifier and the preposition (e.g., between *some* and *of*). Table 9 presents the results of these analyses.

Table 9
ToBI codes for the critical utterances from Experiment 1 with the frequency of accents and breaks listed in parentheses

	Verb accent	Gender accent	Quantifier accent	Quantifier break	Noun accent
Two	H* (11), L + H* (5)	!H* (16)	!H* (12), H* (4)	1-brk (8), 0-brk (8)	!H* (15), L + H* (1)
Some	H* (9), L + H* (7)	!H* (16)	!H* (13), H* (3)	1-brk (15), 0-brk (1)	!H* (16)
Three	H* (10), L + H* (6)	!H* (16)	!H* (13), H* (3)	1-brk (5), 0-brk (11)	!H* (15), L + H* (1)
All	H* (11), L + H* (5)	!H* (16)	!H* (14), H* (2)	1-brk (12), 0-brk (4)	!H* (16)

The relation between discourse status and pitch accent is an active and evolving area of research, characterized by uncertainty about whether the distinctions on either side of the mapping are categorical or continuous (Ladd, 2008; Watson, 2008). However, most theorists posit some kind of ordinal scale of accent prominence which is mapped to an ordinal scale of discourse prominence or predictability (see e.g., Baumann, 2005; Pierrehumbert & Hirschberg, 1990). Such claims are supported by evidence from online comprehension (Dahan, Tanenhaus, & Chambers, 2002; Ito & Speer, 2008; Watson, Tanenhaus, & Gunlogson, in press) and production (Ito, Speer, & Beckman, 2004; Watson, 2008). The L + H* accent is most acoustically prominent and marks contrast or emphasis. The H* accent has been associated with information that is new to the discourse, but it can also be employed when old information is presented in a new light. In the ToBI coding system, the !H* code is used to indicate a high accent which is lower (and less prominent) than a preceding H* accent. It has been associated with information that is discourse accessible but not necessarily in focus.[7]

If we accept this broad characterization, then the accent patterns in these utterances are suitable for the discourse context in which they appeared. In all utterances there was a prominent accent on the verb, consistent with the fact that the action was new information which had not been present in the story. The two nouns were typically produced with a downstepped accent (!H*) suitable for referents that had been mentioned in the story but were not currently in focus. Most of the quantifiers also had a downstepped accent but a few were coded as having an H* accent. This could reflect a subtle difference in the discourse status of the quantifiers; they had not been mentioned in the story but were inferable from the visual context.[8] Note however, that these H* accents were fairly subtle and the quantifier was no longer or louder than in these utterances than in the others.

While the accent pattern was similar across the four types of utterances the break index between the quantifier and the preposition varied. For *some* and *all* the index was typically a 1, indicating that the two words were pronounced as separate words but with no prosodic phrase boundary between them. For *two* and *three* however, the break index was often coded as 0. This could reflect a prosodic difference between the stimuli or it could simply reflect phonological differences in the quantifiers. Both of the numbers end in vowels, thus eliminating a critical cue for identifying a word level break within a prosodic phrase.

B.2. Rating study

B.2.1. Method

Sixteen undergraduate students at Harvard University participated in this study. They received either course credit or $5 for their participation. All participants were native monolingual English speakers.

[7] Some theorists make a binary distinction between *accented* and *deaccented* words. When this distinction is explicitly mapped to ToBI pitch accents, accented items are typically realized with H* and L + H* accents and deaccented items often have !H* accents, rather than no pitch accent at all (see for example, Dahan et al., 2002).

[8] Alternately, the smaller number of downstepped accents on the quantifier could reflect a limitation of either speech perception or speech production. A downstepped accent must be perceived as lower than the high pitch accent that immediately precedes it (Brugos, Shattuck-Hufnagel, & Vielleux, 2006). In most cases the accent on the quantifier was preceded by another downstepped accent (on the gender cue). Thus to be coded as an !H* it would have to lower than this previous downstepped accent, but not at the bottom of the speaker's pitch range.

Table 10
Participants' naturalness ratings for critical utterances

	Two	Some	Three	All
Original	5.6	5.6	6.0	5.6
Gender stress	2.6	4.0	3.1	3.5
Quantifier stress	4.4	3.2	4.6	4.3

This procedure was adopted from a pronunciation acceptability task developed by Kjelgaard and Speer (1999). Participants were asked to judge the naturalness of the pronunciation of a target utterance given the visual scene. They were instructed to rate these items on a scale that ranged from "1" being very unnatural to "7" being very natural. For every trial, participants first saw the names of the featured objects (e.g., "sock and soccer balls"), the target sentence presented orthographically (e.g., "Point to the girl that has some of the socks"), and the corresponding visual display. This information was presented to minimize the effects of online processing on the ratings. Participants then heard the instructions presented auditorily through an adjacent speaker. They had the opportunity to listen to the sentence as many times as they wished before making their judgment.

Participants were presented with 16 utterances. Half of these utterances and scenes came directly from the original study. Four versions of each base item were used to create eight presentation lists such that each list contained two items in each condition and that each base item appeared just once in every list. The other half of the sentences were newly recorded items with focal stress on either the gender cue (e.g., "Point to the girl/boy that has…") or quantifier (e.g., "….two/some/three/all of the socks"). Subsequent ToBI analyses indicated that the focal stress in these utterances was realized as $L + H^*$ accent, which is typically associated with contrast. To ensure that any differences in ratings were linked to the prosodic manipulation, these utterances were carefully matched in other perceivable aspects like speaker identity, utterance length, and shared phonological onset of the featured objects (e.g., *bears* and *berries*). The corresponding visual scenes depicted the distribution of the featured objects across two girl–boy pairs. For the original utterances, these displays were the ones used for the eye-tracking experiment (see Fig. 2). For the prosodically stressed utterances, new displays that featured the same configurations were created. The presentation of sentences was randomized.

B.3. Results

As Table 10 illustrates, the utterances which were used in the experiment were perceived as being fairly natural with a mean rating of 5.7. We found no reliable differences across the four quantifiers ($p > .60$). In contrast, the utterances in which the gender cue was stressed were rated as moderately infelicitous (3.2) while those in which the quantifier was stressed were close to the midpoint of the scale (4.1).

Thus, we conclude that the utterances that were used were natural given the contexts in which they appeared and preferable to utterances with contrastive stress on either the gender cue or the quantifier.

Appendix C. Stimuli items for Experiment 3

Item	Instruction	Distracter
1	Point to the girl that has *some/none/all* of the matches	Maps
2	Point to the boy that has *some/none/all* of the turtles	Turkeys
3	Point to the girl that has *some/none/all* of the sandals	Sandwiches
4	Point to the boy that has *some/none/all* of the papers	Paints
5	Point to the girl that has *some/none/all* of the rats	Rabbits
6	Point to the girl that has *some/none/all* of the pills	Pillows
7	Point to the boy that has *some/none/all* of the cards	Cars

(*continued on next page*)

Appendix C. (*continued*)

Item	Instruction	Distracter
8	Point to the girl that has *some/none/all* of the watermelons	Waffles
9	Point to the boy that has *some/none/all* of the seals	Seagulls
10	Point to the girl that has *some/none/all* of the dogs	Dolls
11	Point to the boy that has *some/none/all* of the mushroom	Muffins
12	Point to the boy that has *some/none/all* of the peas	Pizzas
13	Point to the boy that has *some/none/all* of the bees	Beetles
14	Point to the girl that has *some/none/all* of the baskets	Bats
15	Point to the girl that has *some/none/all* of the socks	Soccer balls
16	Point to the boy that has *some/none/all* of the robes	Roses

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*, 419–439.

Altmann, G., & Kamide, Y. (2004). Now you see it, now you don't: Mediating the mapping between language and the visual world. In J. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action* (pp. 347–386). New York: Psychology Press.

Arnold, J. E., Eisenband, J. G., Brown-Schmidt, S., & Trueswell, J. C. (2000). The rapid use of gender information: Evidence of the time course of pronoun resolution from eye-tracking. *Cognition, 76*, B13–B26.

Baumann, S. (2005). Degrees of givenness and their prosodic marking. In *Paper presented at the international symposium on "Discourse and Prosody as a complex interface"*, September 2005, Aix-en-Provence.

Beckman, M. E., & Hirschberg, J. (1994). *The ToBI annotation conventions*. Columbus, OH: Ohio State University.

Berg, J. (2002). Is semantics still possible? *Journal of Pragmatics, 34*, 349–359.

Bezuidenhout, A., & Cutting, J. C. (2002). Literal meaning, minimal propositions, and pragmatic processing. *Journal of Pragmatics, 34*, 433–456.

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

Breheny, R. (2004). Some scalar implicatures aren't quantity implicatures—but some are. In *Proceedings of the 9th annual meeting of the Gesellschaft fur Semantik* (Sinn und Bedeutung VIIII), November 2004. University of Nijmegen.

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

Brugos, A., Shattuck-Hufnagel, S., & Vielleux, N. (2006). Transcribing prosodic structure of spoken utterances with ToBI., MIT Open Courseware. Available from: http://ocw.mit.edu/OcwWeb/Electrical-Engineering-and-Computer-Science/6-911January–IAP—2006.

Carston, R. (1998). Informativeness, relevance and scalar implicature. In R. Carston & S. Uchida (Eds.), *Relevance theory: Applications and implications* (pp. 179–236). Amsterdam: John Benjamins.

Chierchia, G. (2004a). Scalar implicatures, polarity phenomena, and the syntax/pragmatic interface. In A. Belletti (Ed.), *Belletti structures and beyond*. Oxford: Oxford University Press.

Chierchia, G. (2004b). Numerals and formal vs. substantive features of mass and count. In *Paper presented at Linguistic Perspectives on Numerical Expressions*, Utrecht, 10–11 June 2004.

Dahan, D., Magnuson, J., & Tanenhaus, M. (2001). Time course of frequency effects in spoken- word recognition: Evidence from eye movements. *Cognitive Psychology, 42*, 317–367.

Dahan, D., Tanenhaus, M. K., & Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language, 47*, 292–314.

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

Feeney, A., Scafton, S., Duckworth, A., & Handley, S. J. (2004). The story of some: Everyday pragmatic inferences by children and adults. *Canadian Journal of Experimental Psychology, 58*, 121–132.

Fernald, A., Pinto, J., Swingley, D., Weinberg, A., & McRoberts, G. (1998). Rapid gains in speed of verbal processing by infants in the second year. *Psychological Science, 9*, 228–231.

Frisson, S., & Pickering, M. (1999). The processing of metonymy: Evidence from eye movement. *Journal of Experimental Psychology: Learning, Memory and Cognition, 25*, 1366–1383.

Gadzar, G. (1979). *Pragmatics: Implicature, presupposition and logical form*. New York: Academic Press.

Gibbs, R., & Moise, J. (1997). Pragmatics is understanding what is said. *Cognition, 62*, 51–74.

Glucksberg, S., Gildea, P., & Bookin, H. (1982). On understanding non-literal speech: Can people ignore metaphors? *Journal of Verbal Language and Verbal Behavior, 1*, 85–96.

Grice, H. P. (1957). Meaning. *Philosophical Review, 66*, 377–388.

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

Grodner, D., & Sedivy, J. (in press). The effect of speaker-specific information on pragmatic inferences. In N. Pearlmutter, & E. Gibson (Eds.), *The processing and acquisition of reference*. Cambridge, MA: MIT Press.
Hamilton, W. (1860). *Lectures on logic* (Vol. 1). Edinburgh: Blackwood.
Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on referential domains. *Journal of Memory and Language, 49*, 43–61.
Hirschberg J. (1985). *A theory of scalar implicature*. Doctoral Dissertation, University of Pennsylvania.
Horn, L. (1972). *On the semantic properties of the logical operators in English*. Doctoral Dissertation, UCLA, Los Angeles, CA. Distributed by IULC, Indiana University, Bloomington, IN.
Horn, L. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (Ed.), *Meaning, form, and use in context: Linguistic applications* (pp. 11–42). Washington, DC: Georgetown University Press.
Horn, L. (1989). *A natural history of negation*. Chicago, IL: University of Chicago Press.
Horn, L. (1992). The said and the unsaid. In C. Barker, & D. Dowty (Eds.), *Proceedings of semantics and linguistic theory II* (pp. 163–192). Columbus, OH: Department of Linguistics, Ohio State University.
Ito, K., Speer, S. R., & Beckman, M. (2004). Informational status and pitch accent distribution in spontaneous dialogues in English. In *Proceedings of the international conference on speech prosody*, Nara, Japan.
Ito, K., & Speer, S. (2008). Anticipatory effects of intonation: Eye movements during instructed visual search. *Journal of Memory and Language, 58*(2), 541–573.
Kjelgaard, M., & Speer, S. (1999). Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity. *Journal of Memory and Language, 40*, 153–194.
Koenig, J. (1991). Scalar predicates and negation: Punctual semantics and interval interpretations. In *Chicago Linguistic Society 27*, part 2: Parasession on negation (pp. 140–155).
Ladd, D. R. (2008). Can gradient phonetic distinctions carry categorical pragmatic ones? In *Paper presented at the experimental and theoretical advances in prosody conference*, April 2008, Ithaca, NY.
Levinson, S. (1983). *Pragmatics*. Cambridge: Cambridge University Press.
Levinson, S. (2000). *Presumptive meanings*. Cambridge, MA: MIT Press.
Matin, E., Shao, K. C., & Boff, K. R. (1993). Saccadic overhead: Information processing time with and without saccades. *Perception & Psychophysics, 53*(4), 372–380.
Mirman, D., Magnuson, J. S , Strauss, T. J., & Dixon, J. A. (2008). Effects of global context on homophone ambiguity resolution. In *Proceedings of the 30th annual meeting of the cognitive science society*, July 2008, Washington, DC.
Nadig, A., & Sedivy, J. (2002). Evidence of perspective taking constraints in children's on-line reference resolution. *Psychological Science, 13*, 329–336.
Nicolle, S., & Clark, B. (1999). Experimental pragmatics and what is said: A response to Gibbs and Moise. *Cognition, 69*, 337–354.
Noveck, I. A. (2001). When children are more logical than adults: Experimental investigation of scalar implicatures. *Cognition, 78*, 165–188.
Noveck, I. A., & Posada, A. (2003). Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language, 85*, 203–210.
Papafragou, A., & Musolino, J. (2003). Scalar implicatures: Experiments at the semantics- pragmatics interface. *Cognition, 86*, 253–282.
Papafragou, A., & Tantalou, N. (2004). Children's computation of implicatures. *Language Acquisition, 12*, 71–82.
Pierrehumbert, J. B., & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. R. Cohen, J. Morgan, & M. E. Pollack (Eds.), *Intentions in communication and discourse* (pp. 271–311). MIT Press: Cambridge.
Rayner, K., & Duffy, S. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory and Cognition, 14*, 191–201.
Recanati, F. (2003). *Literal meaning*. Cambridge: Cambridge University Press.
Rips, L. J. (1975). Quantification and semantic memory. *Cognitive Psychology, 7*(3), 307–340.
Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effect of informativity expectations. *Journal of Psycholinguistic Research, 32*(1), 3–23.
Sedivy, J., Tanenhaus, M., Chambers, C., & Carlson, G. (1999). Achieving incremental semantic interpretations through contextual representation. *Cognition, 71*, 109–147.
Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology, 49*, 238–299.
Sperber, D., & Wilson, D. (1986/1995). *Relevance: Communication and cognition*. Oxford: Blackwell.
Swingley, D., & Fernald, A. (2002). Recognition of words referring to present and absent objects by 24-month-olds. *Journal of Memory and Language, 46*, 39–56.
Swinney, D. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior, 18*, 645–660.
Swinney, D., & Prather, P. (1989). On the comprehension of lexical ambiguity by young children: Investigations into the development of mental modularity. In D. Gorfein (Ed.), *Resolving semantic ambiguity*. New York: Springer-Verlag.
Tanenhaus, M. K., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*, 1632.
The British National Corpus, version 3. (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Available from: http://www.natcorp.ox.ac.uk/.
Watson, D. (2008). The role of production factors in acoustic prominence. In *Paper presented at the experimental and theoretical advances in prosody conference*, April 2008, Ithaca, NY.
Watson, D. G., Tanenhaus, M. K., & Gunlogson, C. A. (in press). Interpreting pitch accents in on-line comprehension: H^* vs. $L + H^*$. *Cognitive Science*.