Cognitive Psychology 102 (2018) 105–126

Contents lists available at ScienceDirect

Cognitive Psychology

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

Some inferences still take time: Prosody, predictability, and the speed of scalar implicatures

Yi Ting Huang[a,*], Jesse Snedeker[b]

[a] Department of Hearing and Speech Sciences, University of Maryland College Park, United States
[b] Department of Psychology, Harvard University, United States

ARTICLE INFO

Keywords:
Scalar implicatures
Quantifiers
Prosody
Prediction
Semantics
Pragmatics

ABSTRACT

Experimental pragmatics has gained many insights from understanding how people use weak scalar terms (like *some*) to infer that a stronger alternative (like *all*) is false. Early studies found that comprehenders initially interpret *some* without an upper bound, but later results suggest that this inference is sometimes immediate (e.g., Grodner, Klein, Carbary, & Tanenhaus, 2010). The present paper explores whether rapid inferencing depends on the prosody (i.e., *summa* rather than *some of*) or predictability of referring expressions (e.g., consistently using *some* to describe subsets). Eye-tracking experiments examined looks to subsets (2-of-4 socks) and total sets (3-of-3 soccer balls) following *some* and found early preferences for subsets in predictable contexts but not in less predictable contexts (Experiment 1 and 2). In contrast, there was no reliable prosody effect on inferencing. Changes in predictability did not affect judgments of the naturalness of *some*, when a discourse context was available (Experiment 3). However, predictable contexts reduced variability in speakers' descriptions of subsets and total sets (Experiment 4). Together, these results demonstrate that scalar inferences are often delayed during comprehension, but reference restriction is rapid when set descriptions can be formulated beforehand.

1. Introduction

Contemporary theories of language distinguish between the linguistically encoded meaning of an utterance (*semantics*) and how this meaning is enriched by the context, world knowledge, and speaker goals (*pragmatics*). The division between semantics and pragmatics sheds light on the stability and flexibility of language use during communication, but their boundary can often be unclear and counterintuitive. Take for instance, the dialogue in (1):

(1) Reporter: Will you answer our questions during the press conference?
Politician: I will answer some of them.

Here, we interpret the politician's statement to mean that she will answer one or more of the questions posed to her, but that she will certainly not answer all of them. This intuition is so strong that it is tempting to assume that the meaning of *some* necessarily excludes *all*. Yet, exchanges like (2) demonstrate that this is not the case. Unlike the politician, the movie star uses *some* to be compatible with *all*.

* Corresponding author.
E-mail address: ythuang1@umd.edu (Y.T. Huang).

https://doi.org/10.1016/j.cogpsych.2018.01.004
Accepted 30 January 2018
Available online 06 March 2018

(2) Reporter: Did you address some of the rumors in your book?
Movie star: Of course I addressed some of them. In fact, I addressed all of them.

Yet, others uses of *some* are stubbornly resistant to context information. Exchanges like (3) are infelicitous because they lead to a contradiction between *some* and *none*.

(3) Mob Boss: Did you answer any of the officer's questions?
Flunky: ∗Don't worry! I answered some of them. In fact, I answered none of them!

This shifting pattern of interpretation reflects the distinction between semantically encoded meaning and pragmatic enrichment (Gadzar, 1979; Horn, 1972, 1989). The semantic meaning of *some* is lower bounded: It picks out any amount greater than the minimum value on the quantity scale (i.e., any value greater than *none*, or if the plural is used any value greater than *one*). This semantically encoded content cannot be cancelled. In contrast, *some* excludes *all* by way of an enrichment of this basic meaning. This pragmatic inference is based on listeners' expectation that speakers will be as informative as required but not more informative than is required (Grice, 1975). Thus, it is dependent on conversational goals and beliefs about speakers' knowledge. If the politician in (1) had intended to spill every secret, she could have said (4) instead.

(4) Politician: I will answer all of your questions.

Since she did not use this obvious alternative, listeners can infer that there must be questions that she will not address. By adding an upper bound to *some*, this inference (often called a *scalar implicature*) excludes referents that are compatible with the maximum value on the quantity scale (*all*). Critically, since it is distinct from the semantics of *some*, listeners can still make sense of statements when the inference is cancelled or never calculated, as in (2).[1]

Psycholinguistic studies of scalar implicature have focused on how these two meanings emerge during comprehension. The earliest studies measured response times for judgments of underinformative sentences like *Some elephants are mammals* (Bott & Noveck, 2004; De Neys & Schaeken, 2007; Noveck & Posada, 2003; Rips, 1975). Responding false to these statements indicates the listener has made the scalar implicature, while responding true suggests that she has not. Bott and Noveck (2004) found that participants who judged the statements to be false took longer than those who judged them to be true. This suggests that scalar implicatures are not calculated immediately during comprehension, but instead require time to compute (see Bott, Bailey, & Grodner, 2012 for related work using speed-accuracy tradeoff method, and Tomlinson, Bailey, & Bott, 2013 for work using a mouse-tracking paradigm).

In work using the visual-world paradigm, we found further evidence that scalar implicatures are made after semantic analysis is well under way (see Huang & Snedeker, 2009, henceforth HS, and Huang & Snedeker, 2011). Participants were presented with instructions like *Point to the girl that has some of the socks* while their eye movements were measured to displays featuring a girl with a subset of one item (e.g., 2-of-4 socks) and a second girl with a total set of another item (e.g., 3-of-3 soccer balls). Critically, there was a period of potential ambiguity from the onset of the quantifier to the disambiguation of the final noun (e.g., *-ks*) where the semantics of the quantifier was compatible with both characters. If participants rapidly calculate scalar implicatures, this ambiguity could be resolved since only one of the girls has a proper subset of items. However, after the onset of the quantifier, we found that participants looked equally often at both the subset and total set, leading to slower reference resolution for *some* compared to unambiguous terms like *all*, *two*, and *three*. In fact, evidence of a scalar implicature (as indexed by a reliable preference for the subset compared to the total set) did not emerge until 800 ms after quantifier onset. These results demonstrate that under certain conditions, there is a measurable lag between initial semantic processing and the generation of a pragmatic inference.

The delay observed in HS is consistent with most of the existing research, including many studies that are cited as evidence for rapid calculation of scalar implicatures. For example, Breheny et al. (2006) used a reading paradigm where scalar phrases (e.g., *some of his relatives*) were followed by anaphors that referred back to the excluded complement set (e.g., *the rest*). They found that reading times at the anaphor were shorter in contexts which encouraged the implicature, suggesting that the upper-bounding inference had been completed by the time the anaphor was encountered. However, this study introduces approximately 2000 ms between the onset of the scalar term and the appearance of the anaphor, thus these findings are consistent with a theory where semantic analysis of the scalar term occurs prior to the upper-bounding implicature. Extended time lags are also present in Bergen and Grodner's (2012) reading study (roughly 1800–2400 ms) and Nieuwland et al.'s (2010) ERP study (roughly 1300–1700 ms). In fact, studies that

[1] The exact definition of *scalar implicature* depends on your theory of the phenomenon. Under some accounts, the implicature in (1) results from the insertion of an operator which negates alternatives (like *all*) and can be embedded in semantic structure (see Chierchia, 2004; Chierchia, Fox, & Spector, 2012). On these accounts, all the studies that we will be discussing involve the same enrichment process, and thus are all scalar implicatures. However, in other theories, including the classical Gricean account, scalar implicatures are inferences based on entire speech acts and thus embedded implicatures are impossible (see Geurts, 2009; Breheny, Ferguson, & Katsos, 2013). On these accounts, the utterances in the present study – as well as Huang and Snedeker (2009, 2011) and Grodner, Klein, Carbary, & Tanenhaus, 2010 – are not scalar implicatures because they are embedded in a definite description. Nevertheless, we will be calling them scalar implicatures because: (1) this framing is consistent with the prior studies under discussion; (2) embedded (local) implicatures do occur (see Chemla & Spector, 2011), thus we favor theories that explain them; and (3) psycholinguistic studies suggest that processing patterns in definite descriptions are the same as those found in standard, upward-entailing contexts (c.f. Huang & Snedeker, 2009 to Panizza, Chierchia, Huang, & Snedeker, 2009 or Grodner et al., 2010 to Breheny et al., 2013), thus it is parsimonious to treat these as two examples of a single phenomenon. Those who disagree are free to replace the term *scalar implicature* with their preferred alternative throughout (e.g., *the inference formerly known as scalar implicature*).

substantially decrease the time between the scalar trigger and the anaphor (from roughly 2400 to 900 ms) find no facilitation effects at the anaphor (Hartshorne & Snedeker, 2014; Hartshorne, Azar, Snedeker, & Kim, 2015). This suggests that additional time is needed to access the upper bound from the scalar implicature. Note that Breheny et al. (2006) and Bergen and Grodner (2012) also found that reading times for the scalar trigger (*some*) is longer in contexts that support the scalar implicature compared to ones that do not. While this effect is fast, it has not appeared consistently across studies (c.f. Hartshorne et al., 2015; Hartshorne & Snedeker, 2014; Huang & Gordon, 2011; Lewis, 2013; Politzer-Ahles & Fiorentino, 2013). Furthermore, its interpretation is ambiguous. Slower reading time at the trigger may indicate that the processes for generating the scalar implicature can begin early but that the resulting upper bound emerges much later. This explains why facilitation on the anaphor only emerges when it occurs substantially later than the scalar trigger.

Nevertheless, there is also strong evidence that under some circumstances, the pragmatic content of an implicature is available about as quickly as the semantically encoded meaning of a word. To the best of our knowledge, all this evidence comes from experiments using the visual-world paradigm. The present paper focuses on one of these studies, Grodner et al. (2010; henceforth GKCT), which uses a task that is very similar to HS. Critically, the discoveries that we make by exploring this particular finding in depth will also be relevant to understanding the additional visual-world experiments that also find evidence for early implicatures (e.g., Breheny et al., 2013; Degen & Tanenhaus, 2016). We will return to these studies in the General Discussion. As in the HS studies, GKCT presented participants with sets of objects (e.g., balls and balloons) divided among characters who differed in gender. They were then told to select the mentioned character in the quantified noun phrase (e.g., *Click on the girl who has __ the balls*). However, GKCT made two critical changes to the HS method. First, critical instructions were produced with a different prosodic form. The quantifier had a reduced vowel rather than an unreduced vowel (i.e., *summa* instead of *some-of*). Second, HS included trials with number words in addition to scalar quantifiers, but GKCT did not. As a result, any given set was primarily described using a single quantifier (e.g., *nunna*, *summa*, *alla*). Importantly, the findings from GKCT contrasted sharply with HS. Within 200 ms of the onset of *summa,* participants abandoned looks to the total set and shifted gaze to the subset. In fact, reference restriction for these implicature trials was as rapid as unambiguous *alla* and *nunna* trials.

What accounts for the divergence between these two studies and what can it tell us about the processes by which scalar implicatures are made? We consider two broad possibilities.

1.1. Prosody

In GKCT, the motivation for using prosodically reduced quantifiers was to increase the calculation of scalar implicatures by decreasing the ambiguity of *some*. Specifically, GKCT pointed out that *some* can be interpreted as either a scalar quantifier or a weak determiner (Ladusaw, 1994; Postal, 1964). Since the domain of quantification is unclear when *some* is used with a bare plural in (5a), this may make the weak-determiner interpretation more salient. Importantly, scalar implicatures are infrequent in this context, since *some* is not construed as part of the same quantity scale as *all*. In contrast, when *some* is used in a partitive construction in (5b), the domain of quantification is explicitly defined in terms of a salient maximal set (e.g., all of the apples). Thus, this context provides a stronger basis for generating scalar implicatures (see Degen & Tanenhaus, 2015 for experimental confirmation of this distinction).

(5) Ernie: I ate some apples.
b. Ernie: I ate some of the apples.

GKCT reasoned that using phonologically reduced forms (*summa*) might allow participants to detect the presence of a partitive construction at the onset of the first vowel of the quantifier. This early cue could facilitate processing in one of two ways. One possibility is that the scalar implicature may still need to be calculated from the lower-bounded meaning of *some*, but doing so is faster when the inference is more robust for the partitive. This view is hard to reconcile with GKCT's findings that *some* is given an upper bound as fast as *all* is given a lower bound. We would have to assume that this inference is occurring in real time but is so rapid that we cannot detect the lag using current methods. A second possibility is that scalar implicatures are retrieved by default whenever *some* appears in the partitive, perhaps because these inferences are stored in the lexicon (see Levinson (2000) for the original discussion). While previous reading studies demonstrate that partitives alone are insufficient for accessing scalar implicatures (i.e., contextual support is also needed, see Bergen & Grodner, 2012; Breheny et al., 2006), it is possible that visual-world tasks promote immediate retrieval of a lexically stored upper bound when clear phonological cues are also present.

On either explanation, we would expect slower scalar implicature in HS where use of the unreduced form meant that *some* was initially ambiguous between a scalar quantifier and weak determiner. If the partitive facilitates the rapid online calculation of the implicature, this process might be delayed until the phonological disambiguation of the partitive construction (at *of*) or even later if *some* was initially misconstrued as a weak determiner, and this interpretation had to be revised. Similarly, if phonological reduction facilitates retrieval of a form with a lexicalized upper bound, this route would be inaccessible in the HS study.

1.2. Number words

A second critical difference between the two studies was the range of descriptions adopted across trials. In HS, half of the instructions used number words (e.g., *Point to the girl that has two/three of the socks*) and the other half used scalar quantifiers (e.g., *Point to the girl that has some/all of the socks*). As a result, participants in HS could not reliably predict what lexical label would be used for a particular set (e.g., *some* or *two* could describe 2-out-of-4 socks) or what conceptual encoding would be relevant for the subset

(i.e., proportional quantifier or exact numerosity). In contrast, sets in GKCT were identified using a scalar quantifier on 85% of the trials (e.g., *Click on the girl who has nunna/summa/alla the balls*) while the remaining trials used simple definites (e.g., *Click on the girl who has the balloons*). Thus, lexical labels for any given set were partially predictable before utterances began, and only one conceptualization of quantity was ever relevant for the task. To be clear, listeners in GKCT could not predict *which* set would be the target before sentence onset since the probabilities of referring to subsets, total sets, and empty sets were equated across trials. However, an ideal observer could reliably predict how sets in the scene would be described if they were targets: Total sets were always labeled with *alla,* empty sets always with *nunna,* and subsets were labeled with *summa* on the majority of the trials (70.6% *summa the X's,* 29.4% *the X's*).

The presence of predictable stimuli may have allowed participants to encode the display in a manner directly maps *some* to the subset (verbal-encoding hypothesis). For example, prior to the instruction, listeners who view the display might conceive of the girl with a subset of the balls as *the girl who has some of the balls* and the girl with the total set of balloons as *the girl who has all of the balloons*. As the instructions unfold, listeners could then compare the external speech input to their own internal description of a referent, sticking with a referent when there is a match or ruling it out if they diverge. This would allow listeners to reject the total set as a referent for *some*, without having to first retrieve its meaning and then generate an upper-bounded inference. On this account, scalar implicatures are instantaneous because listeners (like speakers) often encode visual contexts in linguistically relevant ways. In contrast, HS featured two equally frequent conceptualizations of each set. These circumstances may have discouraged predictive encoding during comprehension (since it would be less useful) or blocked the application of previously generated descriptions (since a divergence between internal descriptions and external input would not rule out the possibility that they describe the same referent).

However, GKCT and their colleagues have proposed an alternative explanation for how the presence of number words affects the interpretation of *some* (naturalness hypothesis, see Degen & Tanenhaus, 2015, 2016; Grodner et al., 2010). Specifically, they suggest that introducing numbers within a communicative context focuses listeners' attention to the exact numerosity of small sets (e.g., 2-out-of-4 socks construed exclusively as *two socks*). On this account, delays in interpretation occur because participants in HS do not consider *some* to be good descriptions of small sets when numbers offer a more natural alternative (see Manner maxim in Grice, 1975). We will discuss the details of this proposal prior to Experiment 3. For our present purposes, both HS and GKCT predict that delays in interpreting *some* should occur whenever number words are used in the instructions (albeit for different reasons).

The current study seeks to determine whether differences between HS and GCKT are due to prosodic form of the quantifier or the presence of number-word trials. Experiment 1 tests the prosody hypothesis by assessing interpretation of phonologically reduced forms (i.e., *summa*) in a context where the subset (e.g., girl with 2-of-4 socks) was labeled with *both* scalar quantifiers and number words. Experiment 2 pits the prosody hypothesis against the number hypothesis by factorially manipulating both variables in a between-participants design. Both experiments reveal that prosody has no reliable effect on the timing of scalar implicatures, but the presence of number trials does. When they are absent in the study, the upper bound of *some* is available as quickly as the lower bound of *all.* Finally, we use a ratings task (Experiment 3) and a production task (Experiment 4) to explore why number words affect the interpretation of *some.*

2. Experiment 1

This experiment combined the basic paradigm in the HS study with phonologically reduced quantifiers used in GKCT. On the critical trials, participants were asked to *Click on the girl that has summa the socks* while their eye movements were measured in two visual contexts. On 2-referent trials, participants saw subsets paired with total sets (e.g., girl with 3-of-3 soccer balls). Here, the referent of *summa* is semantically ambiguous but pragmatically unambiguous. On 1-referent trials, participants saw subsets paired with empty sets (e.g., girl with 0-of-3 soccer balls). Here, the referent of *summa* can be distinguished by semantics alone. We also included control trials asking for competing sets (i.e., total set or empty set) using unambiguous scalars (i.e., *alla or nunna*) and filler trials asking for subsets, total sets, or empty sets using alternate descriptions (i.e., *two of* for subsets, *three of* for total sets, and *didn't get* for empty sets). Similar to HS, these trials decreased the predictability of set-to-quantifier mappings since each set type was labeled in two ways, each with equal frequency. To avoid the possibility that interpretations are influenced by multiple usages of *some*, we manipulated quantifier type within subjects but display type between subjects (but c.f. HS's Experiment 3 for a fully within-subjects design).

Experiment 1 allows us to explore two explanations for the rapid preference for subsets in GKCT. If scalar implicatures are immediate when phonological cues to the partitive are available early on, then participants should rapidly access the upper bound in the reduced *summa* trials. As a result, reference resolution should be as fast for the 2-referent trials as it is for the 1-referent trials. However, if immediate implicatures in GCKT reflect additional differences from the current study (e.g., predictability of verbal labels), then participants should be delayed in calculating the upper bound of *some,* as they were in HS. This would result in slower reference resolution in the 2-referent *summa* trials compared to the 1-referent *summa* trials.

2.1. Methods

2.1.1. Participants

Forty English-speaking undergraduate students at Harvard University participated in this study. They received either course credit or $5 for their participation.

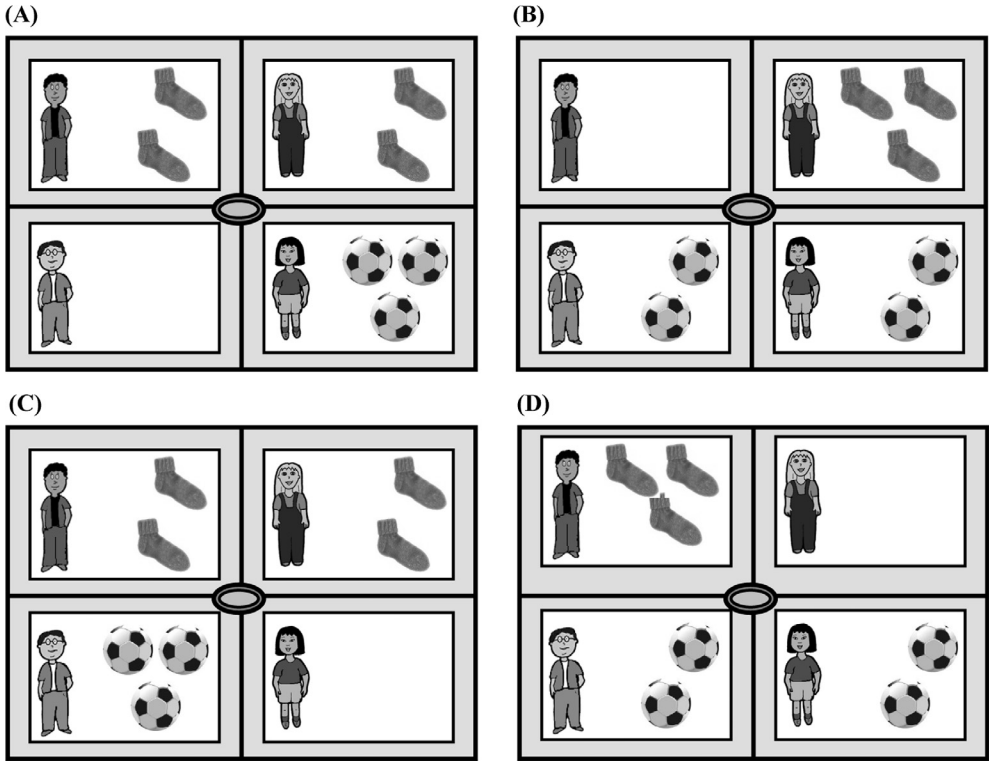

Fig. 1. In Experiment 1, example of 2-referent displays for (A) *summa* trials and (B) *alla* trials and 1-referent displays for (C) *summa* trials and (D) *nunna* trials. Participants here were instructed to *Click on the girl that got ___ of the socks*. The girl with socks was the Target while the girl with soccer balls was the Distractor.

2.1.2. Procedure

Participants sat in front of a computer display and their eye movements to the screen were measured using a Tobii T60 eye-tracker. At the beginning of the study, participants were told that they would hear and see a series of stories about two boys and two girls (i.e., Craig, Pat, Judy, and Cheryl). Every trial consisted of a story followed by a critical utterance. In each story, two sets of objects were distributed and pictures of these objects appeared next the character who received them. The critical utterances instructed participants to select a one of the characters by clicking on it with the mouse. Once the participant did this, the trial ended.

2.1.3. Materials

The stories, displays, and instructions were adapted from Experiment 1 materials in Huang and Snedeker (2009). Participants were assigned to one of two conditions, which varied in terms of the Display Type that was used. In the 1-referent condition, critical trials contrasted a subset (e.g., 2-of-4 socks) with an empty set (e.g., 0-of-3 soccer balls). In the 2-referent condition, they contrasted a subset with a total set (e.g., 3-of-3 soccer balls). All participants heard sentences with different Quantifier Types: Critical *summa* trials (which always picked out the subset) and control trials which picked out the other character using a quantifier. In the 1-referent condition, this alternative was labeled with *nunna.* In the 2-referent condition, it was labeled with *alla*.

Fig. 1 provides examples of the visual displays that were used. Each display depicted four characters who appeared in the same location throughout the study (clockwise 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. On each trial, participants heard a story like (6), in which two types of objects were introduced and distributed among the boy-girl pairs.

(6) 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 (*two socks appear 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 good soccer player, but he thought that Cheryl needed a lot of practice (*nothing appears next to the boy on the lower-left and three soccer balls next to the girl on the lower-right*).

These stories always involved one set of four items that was split evenly between a horizontally adjacent boy-girl pair and another set of three items which was given to one member of the other pair. They established a clear domain of quantification for the critical utterances by introducing the objects as part of a single large set and dividing that set among the characters. For example, after a story like (5), *alla the soccer balls* 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. These stories also ensured that participants knew object labels, which were always referred to with definite noun phrases (e.g., *the socks*) or bare plurals (e.g., *socks*) to avoid priming of the numbers or scalars used in

critical utterances. In off-line judgment tasks, these stories and displays were found to establish the expectations that: (1) quantifiers refer specifically to the sets in the display, (2) objects are identified by basic-level labels, and (3) *summa* is interpreted with a scalar implicature (see Huang & Snedeker 2009, 2011 for more details).

For each story, critical utterances were created like those in (7).

(7) Click on the girl that got summa/nunna/alla the socks.

The gender of the character randomly varied across trials. In the 2-referent condition, a girl was requested when the set of three objects had been given to a girl in the story/display. In the 1-referent condition, a boy was requested when the set of three objects had been given to a girl. The two object types shared a phonological onset (e.g., socks and soccer balls), creating a period of ambiguity when the identity of the noun was uncertain. The requested character is the Target (i.e., the girl with socks) while the other gender match is the Distractor (i.e., the girl with soccer balls in the 2-referent condition or nothing in the 1-referent condition).

Target utterances were recorded by a female actor, and sound files were edited to ensure equal lengths of two windows across conditions: (1) from sentence onset to the gender cue (i.e., *Click on the*) and (2) from the onset of the gender cue to the onset of the quantifier (i.e., *girl that got*). To ensure that sentences were produced in the desired fashion, a trained research assistant coded the *summa* trials using the ToBI annotation system (Beckman & Hirschberg, 1994) and compared them to items from GKCT and HS. The primary difference between GKCT and HS was in the articulation of the consonant at the end of *of*: In HS, it was articulated (e.g., *some of the socks*) while in GKCT it was not (e.g., *summa the socks*). Thus, GKCT featured no word-level break between *summa* and *the* while HS included a consistent break at this juncture. We modelled our utterances on those in GKCT and ensured that the consonant was not articulated. Subsequent ToBI coding confirmed that there was no word-level break. Note that across stimuli sets (HS, GKCT, present study), utterances were always produced without break at the juncture between *some* and *of*. This is because *of* is a clitic that begins in a vowel. Since English is a language that prefers to build right-headed feet, *some-of* makes a perfect metrical foot.

Eight critical items were generated by creating four versions of each base item, which were distributed across four presentation lists such that each list contained four items in each condition and each base item appeared just once in every list. Eight additional filler trials were included and used number words and descriptions. In the 1-referent condition, the Target with the empty set was described as *the girl that didn't get the socks*. In the 2-referent condition, the Target with the total set was described as *three of the socks*. In both cases, the subset was described as *two of the socks*. Because stories and displays preceding the instructions were of the same format for the critical and filler trials, participants could not predict what quantifier would be used to describe a set prior to hearing it. Stimuli for all experiments are provided in Appendices A and B.

2.2. Results

Approximately 2.8% of trials were excluded due to experimenter error or track loss.[2] To isolate changes in interpretation following linguistic cues, we examined fixations to the Target and Distractor across five time windows of the critical utterance:

1. Baseline region: This 500 ms window begins at the onset of the instruction and ends before the onset of the gender cue (i.e., *Click on the*). This region provides a baseline of looks to the display before any gender or quantifier information.
2. Gender region: This 650 ms window begins at the onset of the gender cue and ends before the onset of the quantifier (i.e., *girl that got*). This region provides a direct comparison of looks to the Target and Distractor before any quantifier information. Here, we predict that fixations will shift towards the characters that match the specified gender.
3. Quantifier region: This 700 ms period begins at the onset of the quantifier and ends just before the onset of the disambiguating phoneme (e.g., *summa/nunna/alla the soc-*). In this window, we expect that participants would use information about the quantifier to look to the Target. We expect Target looks to increase when participants hear *nunna* and *alla*, or *summa* in the 1-referent condition. By comparing across trials, we can determine whether reference restriction is relatively delayed when a scalar implicature is required. If implicatures are calculated immediately following an early prosodic cue, we should see a Target preference for *summa* in the 2-referent condition. However, if implicatures are preceded by a period of semantic analysis, then participants should look equally to the Target and Distractor for the *summa* 2-referent condition.
4. Disambiguation region: This 400 ms window begins at the onset of the disambiguating phoneme and ends at the offset of the command (e.g., *-ks*). This region unambiguously resolves the correct referent by picking out his or her objects (e.g., *socks* not soccer *balls*). Here, fixations should be primarily on the Target with relatively few looks the Distractor.
5. End region: This window begins at the end of the sentence and continues for 500 ms. We predict that participants will look at the Target in all trial types.

Each time window begins 200 ms after the relevant marker in the speech stream to account for the time it would typically take to program a saccadic eye movement (Allopenna, Magnuson, & Tanenhaus, 1998; Matin, Shao, & Boff, 1993).

Our primary dependent measure is Target preference, calculated as the number of samples in which the participant looked at the Target for a given trial and window divided by the number of samples in which they looked at the Target or the Distractor (see Huang & Snedeker, 2009, 2011). If participants looked exclusively to the Target, then Target preference in the time window was 1. If they

[2] Raw data and analysis code for all experiments can be found at https://osf.io/94fju.

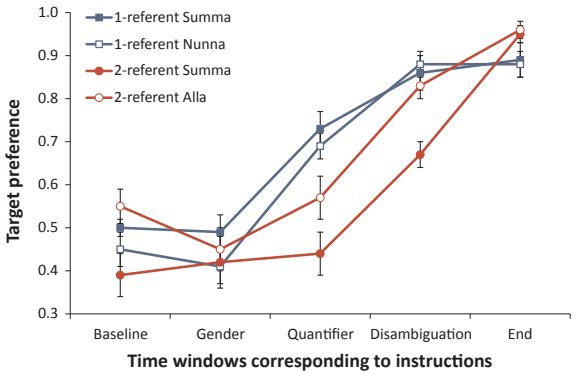

Fig. 2. In Experiment 1, the time-course of Target looks in the 1-referent and 2-referent trials. Error bars represent the standard error of the mean.

looked exclusively to the Distractor, then Target preference was 0. If participants looked at neither object or at both objects equally during a time window, then Target preference was set to 0.5. These data were analyzed in linear mixed-effects models using the lme4 software package in R (Bates et al., 2015). Display Type and Quantifier Type were fixed-effects variables. Across all analyses, we first created maximal models, which included random slopes and intercepts for subjects and items (Barr, Levy, Scheepers, & Tily, 2013). However, in cases where maximal models failed to converge, we adopted simpler models with random intercepts only. Significance tests for fixed effects were estimated via normal approximation of t-statistics. Deviation coding within fixed-effects levels compared condition means to the grand mean.

Prior to quantifier onset, Fig. 2 and Table 1 illustrate that Target looks during the Baseline region were influenced by properties of the Distractor. While Target looks in the 1-referent *nunna* trials were equivalent to the *summa* trials ($\beta = 0.05$, SE = 0.06, t = 0.82, p = .41), Target looks in the 2-referent condition *alla* trials was greater than the *summa* trials ($\beta = 0.17$, SE = 0.07, t = 2.56, p = .01). This led to a significant interaction between Display Type and Quantifier Type ($p < .05$). Since these fixations occur prior to any informative linguistic cue, they likely reflect a visual preference for salient properties of scenes (in this case, unique/larger object sets in *alla* trials). Similar patterns have been found in other visual-world studies (Breheny et al., 2013; Degen & Tanenhaus, 2016; Huang & Snedeker, 2009). Importantly, visual preferences diminished after the onset of the linguistically informative gender cue. During the Gender region, Target fixations converged across conditions, leading to no reliable effects or interactions between Display Type and Quantifier Type (*p's* > .20).

Critically, during the Quantifier region, increases in Target preference emerged in three out of four trial types. In the 1-referent condition, this was true following *nunna* (69%) and *summa* (73%). However, in the 2-referent condition, Target looks increased following *alla* (57%) but not *summa* (44%). This led to significant main effect of Display Type ($p < .001$) and interaction with Quantifier Type ($p < .05$). Planned comparisons revealed Target looks in the 1-referent condition did not reliably differ in *summa* and *nunna* trials ($\beta = 0.05$, SE = 0.04, t = 1.12, $p = .26$). In contrast, Target looks in the 2-referent condition were greater in *alla* compared to *summa* trials ($\beta = 0.13$, SE = 0.06, t = 2.29, $p = .02$). Similar patterns emerged in the Disambiguation region. While Target looks increased for the 2-referent *summa* trials (67%), they continued to lag behind the *alla* (83%), *nunna* (88%), and 1-referent *summa* trials (86%). This led to significant main effects of Display Type ($p < .05$) and Quantifier Type ($p < .01$) as well as an interaction between the two ($p < .05$). Planned comparisons again revealed no difference in Target looks between *summa* and *nunna* trials in the 1-referent condition ($\beta = 0.01$, SE = 0.04, t = 0.34, $p = .73$). In contrast, there remained fewer Target looks for *summa* compared to *alla* trials in the 2-referent condition ($\beta = 0.16$, SE = 0.05, t = 3.37, $p = .001$). Together, this suggests that scalar implicatures are not immediately available during comprehension, even when prosody provides an early signal for the partitive construction.

Finally, after instruction offset during the End region, participants closed in on the Target across all conditions (preference > 85% across trials). Here, Target looks were greater in the 2-referent condition compared to 1-referent condition, leading to a main

Table 1
In Experiment 1, fixed effects (Display × Quantifier Type) in linear mixed-effects models regression model of Target looks during five time windows (Baseline, Gender, Quantifier, Disambiguation, and End).

	Intercept				Display type				Quantifier type				Display × Quantifier			
	β	SE	t	p	β	SE	t	p	β	SE	t	p	β	SE	t	p
Baseline	0.47	0.02	18.74	0.01*	0.01	0.02	0.05	0.96	0.03	0.02	1.52	0.13	0.05	0.02	2.21	0.03*
Gender	0.45	0.02	19.61	0.01*	0.01	0.03	0.45	0.65	0.01	0.03	0.16	0.87	0.03	0.02	1.06	0.29
Quantifier	0.61	0.03	22.31	0.01*	0.10	0.02	4.08	0.01*	0.02	0.03	0.72	0.47	0.04	0.02	1.99	0.05*
Disambiguation	0.81	0.03	28.96	0.01*	0.06	0.03	2.25	0.02*	0.04	0.02	2.84	0.01*	0.04	0.01	2.36	0.02*
End	0.92	0.02	51.03	0.01*	0.03	0.01	2.31	0.02*	0.01	0.02	0.01	0.92	0.01	0.02	0.41	0.68

* $p < .05$ (two tailed).

effect of Display Type ($p < .05$). However, there was no additional effect or interaction with Quantifier Type ($p's > .60$).

2.3. Discussion

In Experiment 1, we found increases in Target looks shortly after the onset of *nunna*, *alla*, and *summa* in a 1-referent context. In contrast, during the 2-referent *summa* trials, participants continued to look equally at the Distractor (total set) and the Target (subset). This suggests that disambiguation is rapid when lexical semantics are sufficient to identity a correct referent. However, when a pragmatic inference is also needed, reference resolution is delayed. These patterns replicate the findings from HS. Critically, they demonstrate that prosody alone cannot account for the differences observed in HS and GKCT.

However, unlike GKCT, Experiment 1 included number trials as filler items. This raises questions about how these trials contribute to delays in interpreting *some.* Recall that numbers may decrease the predictability of set descriptions, forcing listeners to analyze the semantics of *some* before drawing scalar implicatures (verbal-encoding hypothesis). Or they could make *some* unnatural for describing subsets, resulting in processing delays (naturalness hypothesis; Degen & Tanenhaus, 2015, 2016; Grodner et al., 2010). Either way, the goal of Experiment 2 was to reproduce evidence that number words block early access to the upper bound of *some*. To do so, we independently manipulated predictability and prosody in a between-subjects design. Both factors had the possibility of influencing perceptions of the experimental interaction (i.e., calculating likely descriptions, contrasting phonological forms), thus each participant only encountered one level of each factor. Moreover, by having the same speaker produce instructions for all trials, Experiment 2 controlled for additional differences across prior studies (e.g., words in utterances, speech rate, speaker variation). Our primary comparison focused on the interpretation for semantically ambiguous *some* versus unambiguous *all.* Consistent with prior studies, this was manipulated within subjects. Since interpreting *all* is rapid under all accounts, these trials provide a benchmark of the relative delay of deriving scalar implicatures for *some*.

3. Experiment 2

3.1. Methods

3.1.1. Participants

Eighty English-speaking undergraduate students at Harvard University participated in this study. They received course credit or $5 for their participation.

3.1.2. Procedure

The procedure was identical to Experiment 1.

3.1.3. Materials

The eight critical conditions were derived from cells of a $2 \times 2 \times 2$ design. The first factor, Quantifier Strength, was manipulated within-subjects and distinguished the weaker term (*some*) from the stronger one (*all*). The second factor, Prosody, varied whether articulation of the quantifier was phonologically reduced (*summa*) or unreduced (*some-of*). For the sake of clarity, quantifier names in this experiment will refer to variations in the strength of these terms (i.e., some, all), while prosodic variation will be referred to by their phonological differences (i.e., reduced, unreduced). The third factor, Predictability, varied whether sets were labeled in the critical instructions in a uniform fashion or variable fashion. This was manipulated by including fillers that used either scalar terms (*some/all*) to create uniformity, or numbers (*two/three*) to create variability. Experiment 2 also explored two additional factors that might have contributed to differences between Experiment 1 and GKCT findings. Similar to GKCT, we increased the overall tokens of *some* from four to eight trials. We also included filler trials that referred to both 1- and 2-referent sides of the display using scalar quantifiers. Both features may increase the likelihood that participants encode the subset as *some* over the course of the study.

As in Experiment 1, each trial began with a story about objects distributed to characters, followed by an instruction asking for one of the characters. Sixteen critical trials asked for subsets and total sets on the 2-referent side of the display using *some* and *all*. These items were generated by creating eight versions of each base item which were then distributed across eight presentation lists such that each list contained eight items in each condition and that each base item appeared just once in every list. Two types of filler trials were also randomly interspersed. First, 16 filler trials asked for subsets and total sets on the 2-referent side. Depending on the condition, trials used terms from either the quantifier scale (scalar filler: *some/all*) or number scale (number filler: *two/three*). Second, to ensure that all characters were equally likely to be referents in this task, 32 filler trials asked for subsets and empty sets on the 1-referent side. These terms also varied with the condition. In the scalar filler trials, referents were requested using quantifiers (i.e., *some/none*). In the number filler trials, referents were requested using scalars, numbers, and descriptions (i.e., *some/two, none/didn't get*). Critically, the combination of these trials ensured that subsets would be referred to with *some* 100% of the time in the scalar filler condition but only 50% of the time in the number filler condition.

A trained research assistant ToBI coded the critical instructions for the *some* trials across the two levels of Prosody. Similar to Experiment 1, these analyses confirmed there was never a break between *summa* and *the* in the reduced condition (0 in ToBI annotation) but there was a consistent word-level break (1 break) in the unreduced condition. The prosody of the filler trials varied across conditions and matched those of the critical trials.

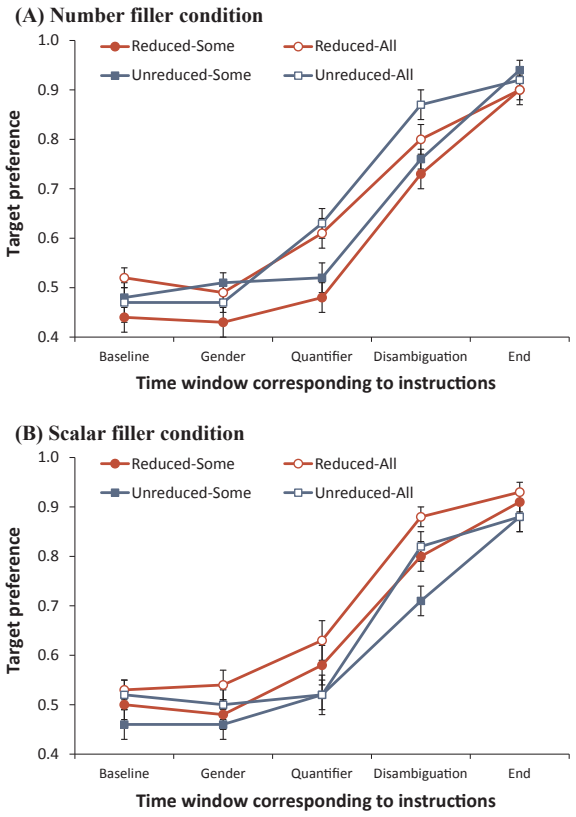

Fig. 3. In Experiment 2, the time-course of Target looks in the (A) number filler condition when verbal labels were variable and (B) scalar filler condition when verbal labels were uniform. Error bars represent the standard error of the mean.

3.2. Results

Approximately 1.1% of trials were excluded due to experimenter error or track loss. Full models first examined how Target preference varied with Predictability, Quantifier Strength, and Prosody across the five time windows: Baseline (700 ms), Gender (600 ms), Quantifier (600 ms), Disambiguation (500 ms), End region (600 ms). Filler type interacted with Quantifier Strength during the Quantifier region ($\beta = 0.02$, SE = 0.01, t = 2.03, $p = .04$) and with Prosody during the Quantifier ($\beta = 0.03$, SE = 0.01, t = 1.96, $p = .05$) and Disambiguation regions ($\beta = 0.03$, SE = 0.01, t = 2.80, $p < .01$). Remaining interactions with Filler were not significant (*p's* > .15). Next, to better understand Predictability effects, we separately analyzed Quantifier Strength and Prosody in the number and scalar filler conditions. During the Baseline region, Fig. 3 and Table 2 illustrates that there were no reliable effects in the scalar filler condition (*p's* > .20). However, consistent with the early visual preference in Experiment 1, Target looks in *all* trials were greater than *some* trials in the number filler condition. This led to a significant main effect of Quantifier Strength ($p < .05$), with no additional effect of or interaction with Prosody (*p's* > .15). Similar to Experiment 1, visual preferences vanished during the Gender region, leading to no reliable effects of Quantifier Strength or Prosody across both filler conditions (*p's* > .15).

Importantly, Target preference differed across conditions in the Quantifier region. Similar to Experiment 1, Target looks in the number filler condition increased following *all* (62%) but not *some* (50%). This led to a main effect of Quantifier Strength ($p < .001$), with no additional effect or interaction with Prosody (*p's* > .50). In contrast, Target looks in the scalar filler condition was similar following *all* (58%) and *some* (55%), leading to no effect of Quantifier Strength ($p > .50$). This suggests that when subsets are consistently labeled with scalar terms, the delays associated with scalar implicatures disappear. Curiously, the scalar filler condition also gave rise to a marginal effect of Prosody ($p < .10$), whereby Target looks were greater for phonologically reduced (61%) compared to unreduced (52%) quantifiers. This pattern is consistent with speakers' tendency to adopt reduced forms when describing predictable referents (Fowler & Housum, 1987; Lam & Watson, 2010). Thus, it is possible that listeners in this study attended to prosodic variation when *some* and *all* were consistently used in the scalar filler condition. Importantly, Prosody did not interact with Quantifier Strength ($p > .50$), suggesting that this effect does not reflect early cuing of partitives for scalar implicatures.

During the Disambiguation region, Target preference for *some* continued to lag behind *all* in the number filler condition (74% vs. 84%). However, this was also true in scalar filler condition (75% vs. 85%), and generated main effects of Quantifier Strength in both conditions (*p's* < .01). This suggests that even while verbal encoding prompts rapid reference restriction for *some*, evidence of co-occurring bottom-up interpretation emerges with the arrival of the disambiguating cue. We will return to this finding in the General Discussion. Consistent scalar descriptions again increased Target preference for reduced (84%) compared to unreduced quantifiers

Table 2
In Experiment 2, fixed effects (Prosody × Quantifier Type) in linear mixed-effects models regression model of Target looks during five time windows (Baseline, Gender, Quantifier, Disambiguation, and End) in the (A) number and (B) scalar filler condition.

(A) Number filler condition

	Intercept				Prosody type				Quantifier type				Prosody × Quantifier			
	β	SE	t	p	β	SE	t	p	β	SE	t	p	β	SE	t	p
Baseline	0.48	0.02	26.16	0.01*	0.01	0.01	0.02	0.98	0.02	0.01	1.98	0.05*	0.02	0.01	1.37	0.17
Gender	0.48	0.02	24.94	0.01*	0.02	0.01	1.06	0.29	0.01	0.02	0.26	0.79	0.02	0.02	1.28	0.20
Quantifier	0.56	0.02	29.42	0.01*	0.01	0.02	0.65	0.51	0.05	0.02	3.40	0.01*	0.01	0.02	0.42	0.67
Disambiguation	0.79	0.02	45.45	0.01*	0.03	0.02	1.52	0.13	0.04	0.01	3.18	0.01*	0.01	0.01	0.87	0.38
End	0.92	0.01	67.13	0.01*	0.02	0.01	1.29	0.20	0.01	0.01	0.21	0.76	0.01	0.01	0.56	0.58

(B) Scalar filler condition

	Intercept				Prosody type				Quantifier type				Prosody × Quantifier			
	β	SE	t	p	β	SE	t	p	β	SE	t	p	β	SE	t	p
Baseline	0.49	0.02	27.67	0.01*	0.01	0.01	0.73	0.46	0.02	0.02	1.18	0.24	0.01	0.02	0.50	0.62
Gender	0.49	0.02	21.71	0.01*	0.01	0.02	0.6	0.55	0.02	0.02	1.30	0.19	0.01	0.01	0.10	0.92
Quantifier	0.56	0.02	24.33	0.01*	0.04	0.02	1.86	0.06	0.01	0.02	0.67	0.50	0.01	0.01	0.65	0.52
Disambiguation	0.80	0.02	31.89	0.01*	0.04	0.02	2.21	0.03*	0.05	0.01	4.16	0.01*	0.01	0.01	0.33	0.74
End	0.90	0.02	52.79	0.01*	0.02	0.02	1.06	0.29	0.01	0.01	0.52	0.60	0.01	0.01	0.19	0.85

* $p < .05$ (two tailed).

(77%), leading to a significant effect of Prosody in the scalar filler condition ($p < .05$) but in not the number filler condition ($p > .10$). Importantly, similar to earlier patterns, there was no interaction between Prosody and Quantifier Strength in either condition (p's $> .30$). Finally, during the End region, participants in all conditions closed in on the Target (preference $> 85\%$ across trials), resulting in no main effects or interactions (p's $> .20$).

Given the extended length of Experiment 2, we investigated when predictability effects emerged in the study by separately analyzing first- and second-half trials. During the Quantifier region, the relative timing of *some* and *all* in the first-half trials varied as a function of filler type. Target preference lagged for *some* compared to *all* in the number filler condition (β = 0.05, SE = 0.02, t = 2.16, $p = .03$), but was equivalent in the scalar filler condition (β = 0.02, SE = 0.02, t = 0.85, $p = .40$). Similarly, second-half trials revealed a reliable effect of Quantifier Strength in the number filler condition (β = 0.06, SE = 0.02, t = 2.68, $p = .01$) but not in the scalar filler condition (β = 0.01, SE = 0.03, t = 0.33, $p = .74$). Thus, parallel predictability effects in both halves indicate that participants anticipated likely set descriptions with strikingly minimal experience. Importantly, analysis of the Disambiguation region adds another wrinkle to this picture. Unlike the Quantifier region, *some* was delayed compared to *all* in first-half trials in both the number (β = 0.04, SE = 0.02, t = 2.61, p = .01) and scalar filler conditions (β = 0.07, SE = 0.02, t = 4.32, p = .001). This suggests that even when verbal encoding promotes rapid access to upper bounds during the Quantifier region, some degree of bottom-up analysis still remains. However, while effects of Quantifier Strength continued in second-half trials of the number filler condition (β = 0.05, SE = 0.02, t = 2.76, p = .01), they disappeared in the scalar filler condition (β = 0.03, SE = 0.02, t = 1.05, p = .29). This suggests that additional experience with predictable contexts can override the last vestiges of bottom-up processing.

Analyzing first- and second-half trials also sheds light on the temporal dynamics of prosody effects. Recall that Prosody and Quantifier Strength never interacted across number and filler conditions (p's $> .30$), yet both were influenced by description predictability. Prosody never influenced interpretation in the number filler condition in first- and second-half trials (p's $> .15$). However, its effects in the scalar filler condition were distinct from patterns found with Quantifier Strength. During the Quantifier region, predictable descriptions promoted effects of Quantifier Strength in first-half trials, but marginal Prosody effects only emerged in second-half trials (first-half trials: $p > .15$; second-half trials: β = 0.05, SE = 0.03, t = 1.90, $p = .06$). During the Disambiguation region, consistent descriptions wiped out effects of Quantifier Strength in second-half trials, but marginal Prosody effects were present across both study halves (first-half trials: β = 0.03, SE = 0.02, t = 1.70, $p = .09$; second-half trials: β = 0.05, SE = 0.03, t = 1.73, $p = .08$). Together, this suggests that consistent descriptions can lead to distinct predictions of lexical labels (i.e., *some* for subsets, *all* for total sets) and prosodic form (i.e., *summa*, *alla*). We will return to this point in the General Discussion.

Finally, we compared *some* trials in Experiment 1 and 2 to understand how scalar implicatures are influenced by the predictability of set descriptions (i.e., whether p(*some*|subset) is 50% vs. 100%) and amount of experience with these descriptions (i.e., encountering 4 vs. 8 tokens of *some*). To control for prosody across experiments, we focused on reduced trials only. During the Quantifier region, Fig. 4 illustrates that 8 predictable trials (Experiment 2, scalar filler condition: 58%) led to significantly more Target looks than 4 unpredictable trials (Experiment 1: 44%; β = 0.14, SE = 0.06, t = 2.40, $p = .02$) and marginally more looks than 8 unpredictable trials (Experiment 2, number filler condition: 48%; β = 0.09, SE = 0.05, t = 1.88, $p = .06$). Yet, within unpredictable trials, there was no effect of trial token (β = 0.05, SE = 0.06, t = 0.80, $p = .40$). This pattern is consistent with the study-halves analyses, and suggests that early reference restriction is influenced by description predictability more so than experience

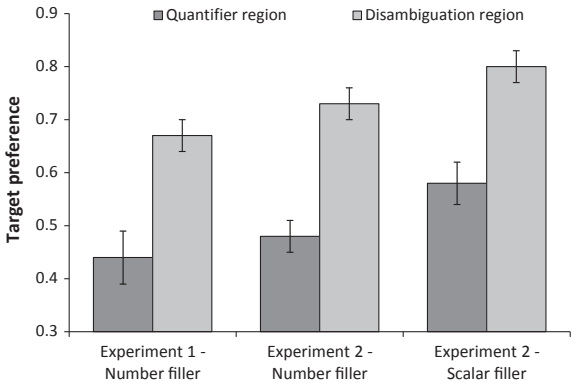

Fig. 4. In Experiments 1 and 2, Target looks during the Quantifier and Disambiguation regions when verbal labels were variable (number filler condition) and when verbal labels were uniform (scalar filler condition). Error bars represent the standard error of the mean.

quantity. Effects of the latter emerged during the Disambiguation region. Target looks for 4 unpredictable trials (Experiment 1: 67%) lagged behind looks for 8 trials, both predictable (Experiment 2, scalar filler condition: 80%; $\beta = 0.15$, SE = 0.05, t = 3.10, $p = .01$) and unpredictable (Experiment 2, number filler condition: 73%; $\beta = 0.08$, SE = 0.05, t = 1.77, $p = .08$). Within Experiment 2, Target looks did not vary with predictability ($\beta = 0.06$, SE = 0.04, t = 1.61, $p = .11$), suggesting that additional experience may be able to speed up bottom-up analysis, even in unpredictable contexts.

3.3. Discussion

In Experiment 2, we found that the timing of *some* interpretation is influenced by how subsets are encoded. When variable labels are used over the course of an experiment, *all* is interpreted more rapidly than *some.* However, when sets are consistently labeled with scalar quantifiers, *some* is interpreted as quickly as *all* with remarkably minimal experience. Comparisons across experiments confirmed that predictable descriptions speed up the timing of *some*, independently of the amount of description experience. In contrast, while reduced forms also facilitated reference restriction, prosody effects emerged after *all* and *some* and in predictable contexts only. Thus, we have a clear answer to the question we posed in the Introduction: Differences between HS and GKCT findings are attributable to the range of descriptions used in the two studies. By recruiting contexts that featured only scalar quantifiers, GKCT created conditions under which the upper bound of *some* restricted reference as rapidly as the lower bound of *all.*

However, as we noted in the Introduction, filler effects lend themselves to two kinds of explanations. Under the verbal-encoding hypothesis, listeners conceptually encode sets in the same way as speakers when likely descriptions can be anticipated. Under the naturalness hypothesis, *some* becomes a less natural description for sets once numbers are also used in utterances. One feature of Experiment 2 seems to uniquely support the verbal-encoding hypothesis. In the scalar filler condition, Target looks were similar after *some* and *all* in the Quantifier region, but delayed for *some* relative to *all* in the Disambiguation region. This demonstrates that rapid predictions of set descriptions can co-occur with a slower process of accessing upper bounds through bottom-up analysis. Early fixations to the Target may reflect verbal encoding (resulting in no difference between *some* and *all*), while later fixations may reflect the consequence of semantic analysis, which concurrently calculates meaning from speech inputs just in case early predictions are incorrect (resulting in slower disambiguation for *some* than *all*). Comparisons of study halves and experiments lend additional support to this interpretation. Late-emerging delays with *some* diminish with more trial tokens, suggesting that listeners become more confident in their predictions as they gain additional evidence of likely set descriptions.

To the best of our knowledge, empirical support for the naturalness hypothesis comes from two sources. First, GKCT reports a ratings study where participants saw displays from the eye-tracking study and judged the naturalness of critical sentences. Half the participants read quantifier sentences only (i.e., *some, all, none*), while the other half read number sentences as well (i.e., *two*, *three*). In the number condition, *two* was judged to be a more natural description of the subset than *some.* Similarly, while *none* and *all* were marginally more natural when numbers were present, *some* was marginally more natural when numbers were absent (p's = .09). Second, Degen and Tanenhaus (2015) elicited naturalness judgments over a wide range of set sizes. While number descriptions were always considered felicitous, the naturalness of *some* varied with the set size. *Some* was highly natural in the middle range (3–7 gumballs) but less so in lower (0–2 gumballs) and higher ranges (8–13 gumballs). Importantly, similar to the GKCT findings, adding number descriptions made *some* an even less felicitous description for small set sizes (1 and 2 gumballs). Together, this suggests that delays with interpreting *some* may reflect a preference for numerical descriptions of subsets, particularly those in the subitizing range. A dominant construal in terms of exact numerosity may interfere with comprehending sentences that refer to sets in terms of their proportional quantity (i.e., *some of the socks*).

Yet, current evidence for the naturalness hypothesis is limited in two ways. First, it fails to explain why the presence of number trials does not slow *some* interpretation when the Distractor is an empty set (i.e., 1-referent trials in Experiment 1). If *some* is an unnatural description whenever numbers are present, it is puzzling why reference resolution was rapid in this context. Second, it is not clear that GKCT's ratings reflect the naturalness of descriptions in HS or the current studies. In GKCT, each trial began with a

numerical description of objects as they appeared in the scene (i.e., *There are four balls, four planets, and four balloons*). These objects were then divided among the characters with no verbal explanation, and the critical instruction presented (i.e., *Click on the girl who has summa the balls*). In contrast, HS and the current study always introduced objects with a story describing why these sets were present and how they were distributed among characters, see (6) for example. Since these richer discourse contexts highlight the division of relevant sets, they may increase the naturalness of scalar descriptions (see also evidence from speaker production in Huang & Arnold, in press).

To explore this possibility, we collected naturalness ratings on our materials and examined how discourse context affects these judgments. Experiment 3 manipulated predictability (scalar trials only vs. scalar/number trials mixed) and discourse context (with stories vs. without stories) in a between-subjects design. Importantly, since the naturalness hypothesis provides an alternative explanation for why interpreting *some* is sometimes slower than *all* (Degen & Tanenhaus, 2015, 2016; Grodner et al., 2010), it predicts two ways in which *some* and *all* ratings will be differently impacted by the presence of number words. First, the naturalness of describing subsets with *some* will decrease when numbers are used, but the naturalness of describing total sets with *all* will not. Second, *two* will be rated as a more natural description of subsets than *some,* but *three* will not be rated as a more natural description of total sets than *all.* Together, these judgment patterns would suggest that specific delays for *some* reflect the degree to which alternative numerical descriptions are preferred for subsets. If, however, the slowness of *some* reflects the initial unavailability of scalar implicatures during comprehension, then off-line judgments of naturalness may not necessarily mirror patterns of on-line reference restriction since these behaviors reflect distinct computations at very different time scales.

4. Experiment 3

4.1. Methods

4.1.1. Participants

Sixty-four English-speaking undergraduate students at the University of North Carolina at Chapel Hill participated in this study. They received course credit or $10 for their participation.

4.1.2. Procedure

The procedure was adopted from the naturalness ratings task described in Appendix B of GKCT with minor modifications. At the beginning of the study, participants were told that they would see displays depicting two pairs of characters and the objects they received. For half the participants, these displays were accompanied with stories describing the scene and for the other half, no stories were provided. Afterwards, one of the characters in the display was highlighted with a red box and referred to with an instruction. GKCT presented these commands in written form. However, to make this task more similar to our eye-tracking experiments, we presented our commands aurally. Participants were told to rate how naturally the utterance identified the character, using a scale that ranged from 1 for very unnatural to 7 for very natural.

4.1.3. Materials

The stories, displays, and sentences were adapted from Experiment 1. Four conditions represented cells of a 2 × 2 between-subjects design. The first factor, Predictability, varied whether the critical sentences labeled sets in a uniform fashion using scalar fillers (i.e., *some/all*) or in a variable manner using number fillers (i.e., *two/three*). The second factor, Discourse Context, indicated whether a story was presented with each display or not. In the story condition, the stories and displays unfolded in the same way as in the eye-tracking studies. In the no story condition, characters and objects were presented in their final configuration. Predictability and Discourse Context were manipulated between subjects since both factors would influence participants' perceptions of the experimental interaction as a whole. Within these four conditions, two additional variables were manipulated. Trial Type distinguished the critical *some* and *all* trials from filler trials in the Predictability manipulation (i.e., *some/all* or *two/three*). Quantifier Strength distinguished the weaker term (i.e., *some*) from the stronger one (i.e., *all*). Both were manipulated within subjects and counterbalanced across four presentation lists.

4.2. Results and discussion

Our dependent measure was the mean naturalness rating of sentences. Fig. 5 illustrates that average ratings always exceeded the midpoint, suggesting that descriptions were generally felicitous. To investigate patterns across conditions, we analyzed the natural log of the ratings using linear mixed-effects models (see Experiment 1 for details on general analytic strategy). Our primary analyses examined Predictability, Discourse Context, and Quantifier Strength as fixed-effects variables and focused on critical trials only. Follow-up analyses directly compared across Scale Type (e.g., *some* versus *two, all* versus *three*) and thus included filler trials as well. To examine whether the naturalness hypothesis holds up in contexts with minimal discourse support, we focused on judgments in the no story condition (Fig. 5a). Recall that with a limited discourse context, GKCT found that number fillers marginally increased the naturalness of *all* but marginally decreased the naturalness of *some.* Our analyses revealed no effects of Quantifier Strength, Predictability, or interaction between them (p's $> .30$). Planned comparisons focusing on the naturalness *some* also revealed no difference in the presence or absence of numbers (5.3 vs. 5.9; $\beta = 0.05$, SE = 0.04, $t = 1.20$, $p > .20$). However, consistent with GKCT's finding and the naturalness hypothesis, we found that *two* was a more natural description of subsets compared to *some* in the number filler trials (6.3 vs. 5.3; $\beta = 0.10$, SE = 0.05, $t = 2.08$, $p = .04$).

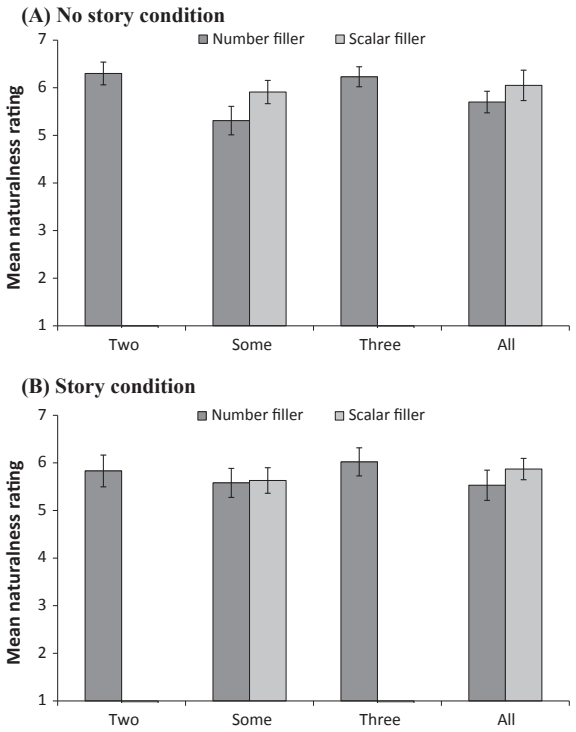

Fig. 5. In Experiment 3, the mean naturalness ratings of utterances describing sets using *two*, *some*, *three*, and *all* in the (A) trials presented without the story and (B) trials with the story, parallel to those in Experiments 1 and 2. Error bars represent the standard error of the mean.

Next, to examine whether the same patterns emerged with a rich discourse context, we repeated our analyses with the story condition (Fig. 5b). These materials approximate those used in HS and the current experiments, thus these analyses address whether naturalness alone can explain the observed delays with interpreting *some*. Similar to the no story condition, there were no main effects or interaction across Quantifier Strength and Predictability (*p's* > .20). Likewise, *some* ratings were unaffected by the presence or absence of numbers (5.6 vs. 5.6; β = 0.01, SE = 0.04, t = 0.16, p > .80). Importantly, unlike the no story condition, *some* was now considered as natural as *two* in the number filler trials (5.6 vs. 5.8; β = 0.02, SE = 0.04, t = 0.38, p > .70). This suggests that a richer discourse context enables subsets to be encoded in terms of either proportional quantity or exact numerosity. This demonstrates that the naturalness hypothesis does not hold up under the conditions in which quantifiers were interpreted in the current study.

Finally, to examine how the naturalness of descriptions changed across contexts, we compared ratings for sentences in the number filler condition. For total sets, *three* was numerically more natural than *all* (6.1 vs 5.6), but the effect of Scale Type did not approach significance (β = 0.02, SE = 0.01, t = 1.57, p = .11). There was no additional effect or interaction with Discourse Context (*p's* > .50). For subsets, there was a marginal effect of Scale Type (β = 0.03, SE = 0.01, t = 1.68, p = .09) and a significant interaction with Discourse Context (β = 0.02, SE = 0.01, t = 2.32, p = .02). This pattern reflected the fact that *two* was more natural in the no story condition compared to the story condition (6.3 vs. 5.8, β = 0.06, SE = 0.04, t = 1.54, p = .12) while *some* was more natural in the story condition compared to the no story condition (5.6 vs. 5.3, β = 0.01, SE = 0.03, t = 0.17, p > .80). This suggests a key difference between GKCT and the current study. With a limited discourse context, exact numerosity is indeed a salient dimension for construing small subsets (Degen & Tanenhaus, 2015). Importantly, adding a richer discourse context facilitates flexible construals of subsets in terms of both *two* and *some*.

Note that these judgments are fully consistent with the verbal-encoding hypothesis, which makes only weak predictions about the naturalness of descriptions. To the extent that participants in the scalar filler condition spontaneously encode sets in these terms (i.e., *some* for subsets, *all* for total sets), it is unsurprising that *some* and *all* are rated highly in this context. Moreover, since the verbal-encoding hypothesis suggests that multiple encodings are salient in the number filler condition, it is unsurprising that numbers and scalars both receive high ratings with a richer discourse context. Importantly, the verbal-encoding hypothesis makes clear predictions about how participants will *describe* sets as speakers. Specifically, when referring expressions are consistent within a context (i.e., scalar terms only), participants should produce these predictable descriptions. However, when referring expressions are unpredictable (i.e., alternating between numbers and scalars), they should vacillate between the observed options.

Experiment 4 tests this prediction by eliciting descriptions of subsets and total sets. This experiment occurred in two parts. Similar to the eye-tracking studies, participants were first presented with a series of stories, displays, and set descriptions. We varied the predictability of these descriptions by introducing utterances involving scalar terms only or a mix of scalar and number expressions. Next, participants saw new stories and displays and were asked to generate their own set descriptions. If the verbal-encoding

hypothesis is correct, then participants will consistently produce scalar descriptions when they are predictable but will show no preference for scalars over numbers when descriptions are unpredictable. Critically, since predictability effects involve calculating likely set descriptions, increased scalar usage should be present both for subsets (i.e., *some*) and total sets (i.e., *all*). In contrast, the focus of the naturalness hypothesis is on how the presence of number words decreases the construal of subsets as *some*. If this account is correct, then number fillers should substantially decrease use of scalar description for subsets but they should not impact descriptions for total sets.

5. Experiment 4

5.1. Methods

5.1.1. Participants

Twenty-eight English-speaking undergraduate students at Harvard University and the University of North Carolina at Chapel Hill participated in this study. They received course credit or $10 for their participation.

5.1.2. Procedure

Participants sat in front of a computer display. At the beginning of the study, they were told that they would hear a story about pairs of characters. After each story, one of the characters would be highlighted with a red box. Their task was to produce an instruction that would allow an imaginary listener to identify that character in the display. They were told that they would first see several examples of how this could be done. During the familiarization phase, participants saw stories and displays, paired with descriptions for the highlighted character. During the test phase, they were presented with similar stories and displays and were now asked to provide an appropriate written description of their own.

5.1.3. Materials

This study featured a 2 × 2 mixed design. During the familiarization phase, the first factor, Predictability, varied the labels used to describe the Target. Similar to prior experiments, this was manipulated between subjects. In the scalar filler condition, participants heard quantities labeled with scalar terms only (i.e., *some* and *all*). In the number filler condition, they heard an equal mix of scalar and number labels (i.e., *some, all, two, three*). These terms appeared in the same carrier phrase as in Experiment 1, and sentences were written at the bottom of the display (i.e., *Click on the girl/boy that got ___ of the ___*). They were paired with eight stories and displays from the Experiment 1 filler trials. During the test phase, the second factor, Display Type, varied the target set to be described. Half the items highlighted the subset while the other half highlighted the total set. Similar to prior experiments, this was manipulated within subjects and was counterbalanced across two presentation lists.

5.2. Results and discussion

Since target characters could be distinguished on the basis of their gender and objects alone (e.g., *girl that has socks* vs. *girl that has soccer balls*), participants omitted modifiers from their descriptions on roughly one-third of trials. Nevertheless, we focused our analyses on descriptions that did include number words or scalar quantifiers to examine how familiarization context influenced which dimension was salient. In the number filler condition, modified utterances accounted for 62% of descriptions for subsets and 65% for total sets. In the scalar filler condition, they accounted for 62% for subsets and 78% for total sets. There were no reliable differences across Display Type or Predictability (*p's* > .15). Remaining descriptions referred to target sets using definite nouns only, and were excluded from further analyses.

To compare production across conditions, we calculated as our primary dependent measure a value called scalar preference. For each trial, scalar preference was 1 if the description included a scalar quantifier (i.e., *some* or *all*) or 0 if it included a number word (i.e., *two* or *three*). These data were analyzed using logistic mixed-effects models (see Experiment 1 for details on general analytic strategy). Fig. 6 illustrates that participants were more likely to use scalar descriptions in the scalar filler condition compared to the number filler condition, leading to a main effect of Predictability ($\beta = 6.93$, SE = 3.26, z = 2.12, $p = .03$). There was no additional effect or interaction with Display Type (*p's* > .30). This demonstrates that recent experience with set descriptions influences how these sets are construed. When quantities were labeled with numbers and scalars, participants produce descriptions that highlight either exact numerosity or proportional quantity. However, when quantities were labeled with scalars only, participants were more likely to hone in on a single dimension and produce scalar descriptions for subsets and total sets.

Next, we examined whether scalar preference in each condition was reliably greater than chance. Since we focused only on trials where number or scalar modifiers were used, we recruited 50% as the relevant benchmark. Consistent with earlier analyses, scalar preference in the scalar filler condition exceeded chance for both subsets (t = 7.07, $p < .001$) and total sets (t = 7.82, $p < .001$). This suggests that repeated exposure to scalar terms led participants to use these labels as dominant descriptions for sets. In contrast, there was no clear preference when describing subsets or total sets in the number filler condition (*p's* > .40). These findings are problematic for the naturalness hypothesis, which attributes delays for *some* to a preference for construing subsets as *two*. Yet, participants' willingness to use either scalar or number descriptions suggests that exact numerosity is not a more salient conceptualization of subsets in the presence of number words. This suggests that other factors are likely responsible for delayed interpretation of *some* (e.g., the initial unavailability of scalar implicatures).

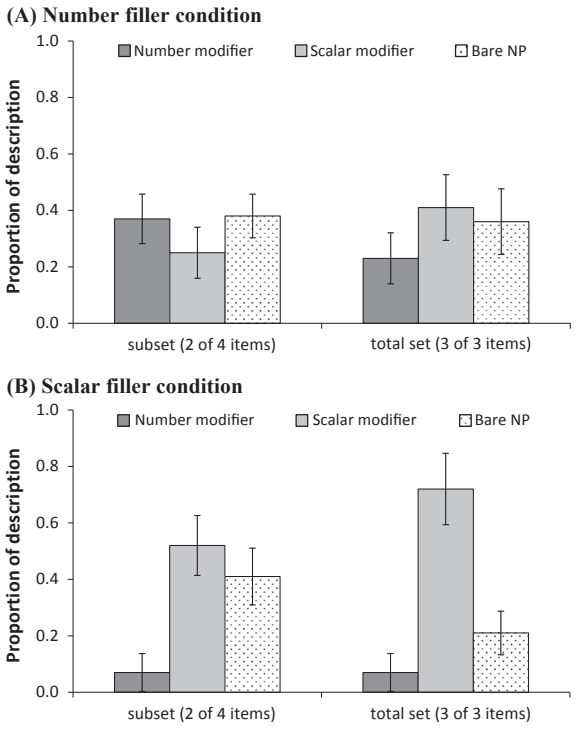

Fig. 6. In Experiment 4, descriptions produced for subsets and total in the (A) number filler condition when familiarized verbal labels were variable and (B) scalar filler condition when familiarized verbal labels were uniform. Error bars represent the standard error of the mean.

6. General discussion

The semantics-pragmatics interface has long been a valuable domain for understanding the representations and computations underlying real-time comprehension. To what extent do words and sentences encode stable meanings across contexts? How do listeners flexibly interpret utterances based on communicative goals? To tackle these questions, the current study examined why routine pragmatic inferences like scalar implicatures are sometimes sluggishly slow (Huang & Snedeker, 2009, 2011; Panizza et al., 2009) and instantaneously fast (Breheny et al., 2013; Degen & Tanenhaus, 2016; Grodner et al., 2010). We first examined whether rapid implicatures are associated with the prosody of *some* or presence of number words. We found no evidence that prosody influences the speed of inferencing, but reference restriction is delayed when descriptions alternate between *some* and *two* (Experiments 1 and 2). We then tested explanations for why number words might have this effect. Contrary to the naturalness hypothesis, we found that *some* and *two* are both sensible subset descriptions with a rich discourse context (Experiment 3). In support of the verbal-encoding hypothesis, we found that consistent descriptions highlight salient dimensions for conceptualizing sets (Experiment 4). Together, these findings suggest pragmatic inferences can arise through two distinct routes during comprehension. When set descriptions are unpredictable, listeners must rely on bottom-up activation from the spoken input to access lexical entries and generate pragmatic inferences. Yet, when set descriptions are predictable, listeners can encode referents in these terms even before they hear the spoken input. These internal, context-specific descriptions can then be compared to external utterances as they unfold and lead to rapid access to pragmatic interpretations. Importantly, what appears to be instantaneous inferencing in fact reflects abilities to construe referents in terms of intended meanings before encountering utterances and map these predictions directly onto the spoken input.

However, there are ways in which these data can be misconstrued. First, one could interpret us as saying that predictable contexts (e.g., ones in which only scalar terms are used) enable experiment-specific strategies, and thus we should only study scalar implicatures in contexts where set descriptions vary from trial to trial. This is not our intention. Effects of predictability emerge early in studies and sustain over time. They may be strategic in the sense of being shaped by experiences with the task at hand. Yet, they are also deployed rapidly and consistently. Thus, we prefer to construe these effects as further evidence that language processing is dynamic and adaptive. In other words, strategies this good deserve some respect. Second, our claim that bottom-up analysis leads to slower scalar implicatures than verbal encoding could be misconstrued as a claim about modularity, namely that there exists a stage of semantic analysis that is wholly independent of pragmatics that must be completed before an implicature can occur. We see no reason to take such a strong position. Incrementality and interactivity appear to be the norm in language comprehension. Instead, our claim is far more modest: Scalar implicatures are computations that take place in real time, and are not the stored products of these computations. Consequently, accessing an upper-bounded interpretation cannot be instantaneous. When scalar implicatures *appear* to be atemporal, this is because the work was done beforehand. While we consistently find a 600–800 ms delay in visual-world tasks, we

remain agnostic as to why implicatures take as long as they do. This may reflect the contextual richness and complexity of these inferences. Finally, stating that the lower-bounded meaning of *some* is often available before its upper bound, in no way implies that pragmatic inferencing is less important, more fragile, or fully independent from semantic decoding. We are simply arguing that the path from sound to inference passes through the lexicon, a claim that we hope is uncontroversial.

In the remainder of our discussion, we will examine how verbal encoding explains additional evidence of rapid scalar implicatures found in visual-world studies (Breheny et al., 2013; Degen & Tanenhaus, 2016). Next, we will discuss what our findings reveal about the underlying mechanisms of predictions during comprehension, and ways in which they support and constrain influential accounts of predictions in language and other domains (Dell & Chang, 2014; Clark, 2013; Pickering & Garrod, 2013). Finally, we will consider the implications of our findings for computational-level questions of communication and why predictions are particularly useful for comprehending rapidly unfolding speech (Gibson, Bergen, & Piantadosi, 2013; Goodman & Frank, 2016; Levy, Bicknell, Slattery, & Rayner, 2009).

6.1. Accounting for additional evidence of rapid implicatures

While the current study focused on resolving differences between HS and GKCT, this section examines how verbal encoding explains additional evidence of rapid implicatures from visual-world experiments by Breheny et al. (2013) and Degen and Tanenhaus (2016). These studies differ in their procedures and materials, but we were interested the degree to which the ingredients for verbal encoding were present. Do familiarization events unfolded slowly over time so that participants could generate appropriate labels for referents? Do visual displays provide salient contrast between quantities? Most importantly, are sets described in a uniform manner across sentences in the study? To evaluate the latter, we used Bayes' theorem to calculate the predictability of subset descriptions in each experiment:

(8) $\mathrm{p}(\textit{some}|\mathrm{subset}) = \frac{\mathrm{p}(\mathrm{subset} \mid \textit{some})\mathrm{p}(\textit{some})}{\mathrm{p}(\mathrm{subset})}$

Across all experiments, Table 3 illustrates that utterances referred to subsets on roughly half of the trials (i.e., p(subset) ranged from 50 to 58%). Thus, variation in the predictability of set descriptions was largely driven the consistency of *some* usage (i.e., p(subset|*some*): how often *some* described non-subsets) and frequency of *some* usage (i.e., p(*some*): how often *some* occurred relative to other descriptions).

Breheny et al. (2013) examined the time course of interpreting *some* and *all* using both quantities of continuous substance (e.g., water) and discrete objects (e.g., socks). For example, on a continuous-substance trial, participants were first familiarized with a transfer event that moved water from two pitchers into two bowls. During the test phase, they heard a sentence like (9) while their eye-movements were measured to a display featuring a bowl with half of the water containing limes (subset) and a bowl with all of the water containing oranges (total set).

(9) The man has poured some of the water with limes into the bowl on tray A and all of the water with oranges into the bowl on tray B.

While critical sentences disambiguated the target bowl at the onset of *limes*, participants shifted their looks to the subset immediately after the offset of *some*. Reference resolution in these trials was similar to *all* and faster than late-*some* trials, which asked for *some* in the presence of two subsets. This demonstrates that listeners accessed an upper bound after *some* and used this to reject the total set. However, closer inspection suggests that study materials provided ideal conditions for verbal encoding. Critical sentences always labeled subsets with *some* and total sets with *all*. Filler sentences referred to sets using other descriptions (e.g., definite expressions like *the staplers*, count phrases like *three forks*), but the infrequency of such trials meant that *some* remained the dominant descriptions for subsets (76%). Comparisons across studies reveals that the likelihood of subsets were labeled with *some* was lower than predictable contexts in the current study (100%), but it was very similar to GKCT (71%). Since familiarization events unfolded

Table 3
In visual-world studies, the speed of interpreting *some* as a function of experimental statistics. P(*some* | subset) refers to how often subsets were described with *some* (i.e., predictability of set descriptions). P(subset | *some*) refers to how often some described non-subsets. P(*some*) refers to how often *some* occurred relative to other descriptions. P(subset) refers to how often subsets were the Target relative to total sets.

Study	Speed of *some*	p(*some* | subset)	p(subset | *some*)	p(*some*)	p(subset)
Current study Exp 1: Two-referent display (16 total trials)	Slower than *all* (Quantifier region)	50%	100%	25%	50%
Current study Exp 2: Number filler (64 total trials)	Slower than all (Quantifier region)	50%	100%	25%	50%
Current study Exp 2: Scalar filler (64 total trials)	As fast as *all* (Quantifier region)	100%	100%	50%	50%
Breheny et al. (2013) (50 total trials)	As fast as *all* (Quantifier offset)	76%	100%	44%	58%
Degen and Tanenhaus (2016) Exp 1: Numbers absent (64 total trials)	As fast as *all* (Quantifier region)	75%	75%	50%	50%
Degen and Tanenhaus (2016) Exp 2: Numbers present (96 total trials)	As fast as *all,* Slower than numbers (Quantifier region)	50%	75%	33%	50%

over 25 s, participants had ample opportunity to generate descriptions on this basis.

More recently, Degen and Tanenhaus (2016) use an innovative "gumball" paradigm to vary the set sizes associated with *some* and *all*. On each trial, participants saw colored gumballs located in an upper chamber (e.g., 8 blue, 4 orange) and clicked on the display to transfer sets into the lower chamber. About 500 ms later, they heard sentences like (10) while their eye-movements were measured to the lower chamber, which contrasted a subset of one color (e.g., 2-out-of-8 blue gumballs) and a total set of another (e.g., 4-out-of-4 orange gumballs).

(10) You got some/all of the blue/orange gumballs.

Degen and Tanenhaus (2016) used this paradigm in two experiments. In Experiment 1, trials included quantifier descriptions only, and target fixations immediately increased after both *some* and *all*. Importantly, the authors argued that speedy reference restriction could not reflect verbal encoding since set sizes varied across trial displays (e.g., subsets could be 2, 3, 4 or 5 out of 8 gumballs). Yet, since Experiment 1 never used number words, exact numerosity was never a relevant task dimension. Thus, listeners could converge on proportional quantity as the relevant encoding after hearing *some* and *all* only. This would be akin to how speakers in the current study rarely described sets with numbers when they were labeled with scalars only (see also discourse manipulations in Huang and Arnold (in press)). Moreover, visual contrast in upper- and lower-gumball chambers easily distinguished proportional quantity, independent of exact numerosity (i.e., seeing color in the upper chamber implies subset, seeing no color implies total set). The authors also argued that verbal encoding was unlikely since Experiment 1 also included "garden-paths" trials, which labeled total sets with *some* and subsets with *all* on 25% of the time. Thus, unlike prior work, reference to subsets was not perfectly predicted by *some* usages. Nevertheless, given the high occurrence of *some* trials (50%), the predictability of describing subsets using *some* remained comparable to past studies (75%).

To decrease the predictability of set labels, Degen and Tanenhaus (2016) conducted Experiment 2, which mixed *some* and *all* with number word trials (i.e., *two*, *three*, *four*, *five*). Adding these trials meant that listeners could no longer reliably predict subset descriptions. Similar to unpredictable contexts in HS, target looks immediately increased following the onset of number words. However, unlike HS and the current study, the time course of target looks was similar for *some* and *all*. Degen and Tanenhaus (2016) also suggested that visual preferences for larger sets may explain why *some* was slower than *all* in HS. In time windows before and following quantifier onset, target looks were greater when *all* referred bigger sets (e.g., 4 or 5 gumballs) and *some* referred smaller sets (e.g., 2 or 3 gumballs). Yet, closer inspection of the data suggests important qualifications to their interpretation. First, when labels were predictable in Experiment 1, similarities in *some* and *all* were driven by increases in target looks immediately after both quantifiers (pg. 184, Fig. 3b). However, when labels were unpredictable in Experiment 2, similarities instead reflected delays in interpreting both quantifiers (pg. 191, Fig. 7b). Thus, while the slowness of *all* is surprising, its similarity with *some* does not imply that scalar implicatures were available early in comprehension. Second, while visual preferences may explain why fixations differ when targets are distinct sets, they fail to address why the timing of *some* changes across identical displays in the current Experiment 2. Recall that interpreting *some* was quick with predictable descriptions but slow with unpredictable ones. These trials involved the same quantifier and set size (i.e., *some* refers to 2-out-of-4 objects), thus variation in timing must reflect its relationship to co-occurring trials (i.e., likelihood that subset are labeled with *some*).

6.2. How do listeners generate predictions during comprehension?

Broadly speaking, the current findings are consistent well-known phenomena within language prediction. Across methods and materials, language comprehension makes pervasive use of message-level information to anticipate upcoming words. In visual-world studies, informative verbs (e.g., *The boy will eat…*) motivate predictive fixations to semantically related objects (e.g., cake) (Altmann & Kamide, 1999; Kamide, Altmann, & Haywood, 2003). In EEG studies, predictable semantic contexts (e.g., *She measured the flour so she could bake a…*) lead to a P130 response after misspelled plausible words (e.g., *ceke*) compared to non-words (e.g., *tont*) (Kim & Lai, 2012). Similar effects emerge in syntax, whereby increased experience within (Fine & Jaeger, 2013; Kowalski & Huang, 2017) and across contexts (Wells, Christiansen, Race, Acheson, & MacDonald, 2009) speeds up interpretation of infrequent and less predictable constructions. Nevertheless, the sheer diversity of phenomena can make isolating underlying mechanisms challenging. In the literature, predictions are sometimes described as products of a single domain-general process for imitation (e.g., Pickering and Garrod (2013)'s Imitation/Forward modeling) or action planning (e.g., Clark (2013)'s Action-oriented prediction) while at other times, they are explained in terms of multiple interactions across domain-specific systems (e.g., Dell and Chang (2014)'s Error-based learning/-chain model). Importantly, by examining language predictions within the single test case of scalar implicatures, the current study may offer unique insights into their cognitive underpinnings.

First, parallels context effects across comprehension (Experiment 2) and production (Experiment 4) are consistent with the notion that systems for speaking are recruited while listening. While it has been argued this overlap arises from listeners' covert imitation of speakers' actions (Pickering & Garrod, 2013), the current study suggests that predictions can reflect a much simpler propensity for implicit naming. It is striking that listeners in our studies converged on likely descriptions with very minimal speaker cues. These effects are reminiscent of spontaneous word-form selection found in tasks involving no spoken input or verbal response (Jescheniak & Levelt, 1994; Meyer, Belke, Telling, & Humphreys, 2007). In a non-linguistic memory task, participants look longer to pictures with 3-syllable names like *elephant* compared to 1-syllable names like *ball* (Zelinsky & Murphy, 2000). Similarly, pictures with high name agreement (e.g., apple is often described as *apple*) are more likely to promote spontaneous retrieval of labels compared to ones with lower name agreement (e.g., couch can be described as *couch*, *sofa*, *loveseat*) (Dikker & Pylkkanen, 2011, 2013). Remarkably, implicit

naming is also demonstrated in 18-month-olds infants, who generate longer looking times to pictures with shared onsets (e.g., *dog-door*) compared to pictures without phonological overlap (e.g., *dog-boat*) (Mani & Plunkett, 2010, 2011). Together, this suggests that procedures for implicit naming are fairly automatic, and do not require an extensive lexicon or metalinguistic awareness.

Second, while predictions are often characterized as a monolithic, domain-general process, our findings suggest that a given context can support multiple expectations of linguistic representations. Recall that while reference restriction for *some* was faster with scalar fillers compared to number ones, the timing of *all* remained unchanged. This suggests that sets were pre-encoded in terms of lexical representations (i.e., subsets linked to *some*, total sets to *all*). If sets were instead encoded with respect to phonological representations (i.e., subsets linked to sʌm, total sets to ɔl), then early access to sound forms would have eliminated semantic analysis for *some* and *all* and increased the speed of both quantifiers. Importantly, evidence of prosody effects suggests that predictable contexts can also increase anticipation of reduced phonological forms (i.e., *summa*, *alla*). It is well established that speakers recruit lower pitch, decreased intensity, and shorter duration to distinguish previously mentioned referents from unmentioned ones (Fowler & Housum, 1987; Lam & Watson, 2010), and listeners predict referents on this basis (Dahan, Tanenhaus, & Chambers, 2002; Huang, Newman, Catalano, & Goupell, 2017). Notably, even when the overall speed of interpretation increased with trial tokens, prosody effects never emerged in number filler (unpredictable) contexts. This absence may reflect the fact that prosody was characterized by duration changes in the current study. This has been argued to be an unreliable cue to speaker meaning since it is confounded with production difficulty (Arnold & Watson, 2015; Watson, 2010). Thus, reduced forms may have been insufficient for triggering referential predictions, except when uniform descriptions drew attention to this prosodic variation.

Finally, the current findings demonstrate that the presence of top-down predictions does not preclude the need for bottom-up analysis. This furthers our understanding of classic evidence of predictions in visual-world studies, which reveals that fixations are immediately sensitive to later-arriving cues, but integrating early predictions with subsequent analysis takes time. This is true in garden-path recovery, whereby fixations linger on objects that match initial misanalyses, even after the arrival of late conflict (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). This is also true when late-arriving evidence conforms to initial predictions (Altmann & Kamide, 1999). While informative verbs (*eat*) increase fixations to likely objects more so than uninformative verbs (*move*), this difference continues for approximately 500 ms after the final noun (*cake*). This pattern is similar to Experiment 2, where reference restriction in the number filler condition is faster following informative *all* relative to less informative *some*, and this difference continues after the disambiguating cue (e.g., *-ks* in *socks*). Given this persistent delay, the scalar filler condition isolates whether top-down predictions rule out the need for bottom-up analysis by creating a context where pre-encoded meanings of *some* are as informative as *all* while semantic meanings remain less informative. If top-down predictions obviate bottom-up analysis, then the timing of *some* and *all* should now be equivalent throughout the entire utterance. Instead, evidence of a late-arriving delay suggests that co-occurring bottom-up analysis generates a lower-bounded interpretation that must be integrated after the disambiguating cue.

The presence of top-down predictions and bottom-up analysis is broadly consistent with the Error-based Learning/P-chain account (Dell & Chang, 2014), which argues that bottom-up interpretation provides a means for assessing the accuracy of predictions. This is critical when listeners have a more limited basis for generating predictions early in communicative interactions, and explains why fixation patterns vary with study halves in the scalar filler condition. Recall that in first-half trials, *some* was as fast as *all* immediately after the quantifier, but *some* was slower than *all* after the disambiguating cue. Even though subsets were twice as likely to be described with *some* in the scalar filler condition compared to the number filler condition, listeners' estimates of these probabilities are inherently noisy when they have limited experience with set descriptions early in the study. Under these circumstances, they benefit from simultaneously retrieving the lower-bounded semantics of *some* as well as anticipating that subsets will likely be described with *some*. However, as experience with set descriptions accumulates, listeners can realize that predicted meanings overlap with those from bottom-up analysis and decrease their reliance on the slower process. This explains why late delays with *some* disappear in second-half trials.

6.3. Why do listeners predict during comprehension?

Beyond describing *how* listeners predict, the current study sheds light on *why* predictions are useful during communication. Our findings demonstrate that predictions are highly sensitive to task goals: What you predict depends on why you are predicting. When utterance interpretation satisfies reference restriction (Experiments 1 and 2), listeners quickly and spontaneously recruit consistent descriptions to predict subsets after hearing *some*. Likewise, when utterance production satisfies reference disambiguation (Experiment 4), speakers borrow from existing set descriptions. Finally, when utterance interpretation informs naturalness judgments, discourse context influences the degree to which exact numerosity and proportional quantity are sensible descriptions (Experiment 3). Together, these findings suggest that sets can be encoded along multiple dimensions, thus listeners must interpret utterances to isolate speakers' perspectives. Importantly, this view is at odds with the naturalness hypothesis, which argues that sets take on default encodings based on perceptual properties such as set size (Degen & Tanenhaus, 2015, 2016; Grodner et al., 2010). This notion is puzzling given the multitude of set descriptions within natural language, e.g., referring to 2-out-of-5 houses in terms of object category (e.g. *There were pretty houses on that street*), membership within a set (e.g., *I saw the houses that you had talked about*), relationship to a larger total set (e.g. *Some of the houses were for sale*). This variation suggests that speaker decisions about set descriptions are inherently tied to dimensions that are relevant for current communicative goals (Huang & Arnold, in press). Since this varies across contexts, listeners benefit from mechanisms that anticipate speaker-specific expressions.

The current findings also clarify the role of formal accounts of pragmatic inferencing like the Rational Speech Act (RSA) model (Goodman & Frank, 2016). Following Grice (1975), the RSA model explains how listeners arrive at sensible interpretations of seemingly underinformative utterances. For example, a speaker may state that *My friend has glasses* in a context where one face is

wearing glasses and another face is wearing glasses and hat. Under the RSA model, listeners realize that referent descriptions (e.g., *glasses*, *hat*) can be ordered based on how much information they convey about the state of the world (*w*). Weaker descriptions are compatible with multiple referents (e.g., P(*w* | *glasses*) while stronger descriptions are compatible with a single referent (e.g., P(*w* | *hat*). Moreover, by assuming that speakers provide both relevant and parsimonious descriptions, listeners can infer that use of weaker descriptions implies that stronger alternatives are not true (i.e., *My friend has glasses* refers to the face wearing only glasses). The same framework also captures why interpreting *some* is faster in predictable contexts. Since utterance informativity is directly related to the probability of inferring *w* (i.e., subset) given the utterance (i.e., *some*), this value is higher when descriptions of *w* are uniform in predictable contexts compared to when they are variable in unpredictable contexts.

At a computational level, the RSA model accurately highlights the importance of pragmatic inferencing during communication. However, at an algorithmic level, it does not naturally explain patterns of production and comprehension in the current study. For example, participants did not appear to assume that speakers should or will provide parsimonious descriptions. Speakers in Experiment 4 often produced numbers and scalar modifiers, despite the fact that definite descriptions were sufficient for disambiguation (e.g., no need to say *girl with some of the socks* when girls had either socks or soccer balls). Likewise, listeners in Experiments 1 and 2 restricted reference using early-arriving quantifiers, despite the fact that final nouns were sufficient for referent identification. Similar effects are found with color modification (Rubio-Fernández, 2016; Sedivy, 2003), where speakers produce overinformative descriptions (e.g., saying *red car* when there is only one car in the scene), and this facilitate reference restriction in listeners (e.g., looks to the car following *red*). Such patterns are puzzling from the perspective of parsimony, but they make sense in light of the environments in which listeners interpret utterances during real-world communication (Gibson et al., 2013; Levy et al., 2009). Since speech streams are subject to random or systematic degradation (e.g., background noise makes it difficult to distinguish cohorts like dogs vs. dolls), speakers can provide redundant information to ensure that their message is received (e.g., *some of the dogs* is unlikely to be misinterpreted as *all of the dolls*). Conversely, since rapidly unfolding speech streams present inherent challenges to bottom-up analysis, listeners can rely on top-down predictions can anticipate meanings via contextual cues (e.g., subsets often labeled with *some*).

One remaining question is whether pre-encoded meanings from top-down predictions are identical to those derived from bottom-up inferencing. While the two may be more or less equivalent for lexical predictions (e.g., anticipating *cake* after *eat* is similar to recognizing *cake* after *cake*), pragmatic inferencing involves assessing utterance meaning (e.g., *some* implies *not all*) as well as speaker goals (e.g., Does she have knowledge of stronger alternatives? Why didn't she unambiguously say *only some* or *not all*?). Prior research demonstrates that listeners are sensitive to speakers' use of scalar implicatures to convey criticism politely, and are more likely to derive inferences in face-boosting (e.g., *Some people loved your poem*) compared to face-threatening contexts (e.g., *Some people hated your poem*) (Bonnefon, Feeney, & Villejoubert, 2009). Similarly, recent work suggests that deriving ironic interpretations is distinct from interpreting opposites, despite equivalent truth conditions (Adler, Novick, & Huang, 2018). Nevertheless, there is also substantial evidence that generalized implicatures are guided by linguistic factors, including semantic entailment (Chierchia, 2004; Chierchia et al., 2012) and salience of stronger alternatives (Rees & Bott, in press; Skordos & Papafragou, 2016). This raises the possibility that inferred interpretations are regularly accessed through linguistic predictions of likely meanings.

7. Conclusions

Understanding why language comprehension is often fast and sometimes slow involves isolating the processes that contribute to interpretation along different time scales. This includes the moment-to-moment changes in comprehension that follow the onset of spoken input but also the expectations that listeners generate prior to this point. By examining the test case of scalar implicature, the current study suggests real-time comprehension involves the dynamic interplay between two mechanisms for interpreting utterances: (1) rapid access to pragmatic interpretations when verbal encoding generates internal, context-specific descriptions before the external input unfolds; (2) a slower process of calculating pragmatic inference from bottom-up input through initial semantic analysis of lexical meaning. Understanding these processes will require further research to determine the conditions under which they occur, the signatures of their use, and the representations and computations that underlie them.

Acknowledgments

This work is supported by grants from NICHD to YH (HD061173) and NSF to JS (BCS-0623845). We thank Noemi Hahn, Kate McCurdy, Elizabeth Casserly, Amanda Worek, and Philip Kim for their help with data collection and ToBI analyses. We also thank Manizeh Khan for comments on an earlier draft and Dan Grodner and Heather Ferguson for sharing the stimuli used in their studies. This work benefited from the comments of audience members at XPRAG 2009 and the XPRAG master class in 2013. Author address: Department of Hearing and Speech Sciences, 0100 Lefrak Hall, College Park, MD 20742. Email address: ythuang1@umd.edu.

Appendix A

Stimuli items for Experiment 1. On critical trials, Targets and Distractors shared an average of 270 ms of overlap in phonological onset.

Type	Item	Sentence	Distractor
Critical trials	1	Click on the girl that got summa/alla the pills	Pillows
	2	Click on the girl that got summa/alla the watermelons	Waffles
	3	Click on the boy that got summa/alla the peas	Pizzas
	4	Click on the boy that got summa/alla the robes	Roses
	5	Click on the girl that got summa/alla the sandals	Sandwiches
	6	Click on the girl that got summa/alla the rats	Rabbits
	7	Click on the boy that got summa/alla the mushrooms	Muffins
	8	Click on the boy that got summa/alla the bees	Beetles
Filler trials	1	Click on the boy that didn't get anything (paints)	Papers
	2	Click on the girl that didn't get anything (socks)	Soccer balls
	3	Click on the boy that got two of the cards	Cars
	4	Click on the girl that got two of the dogs	Dolls
	5	Click on the girl that got two of the turtles	Turkeys
	6	Click on the boy that got three of the matches	Maps
	7	Click on the boy that got two of the baskets	Bats
	8	Click on the girl that got three of the seals	Seagulls

Appendix B

Stimuli items for Experiment 2. On critical trials, Targets and Distractors shared an average of 290 ms of overlap in phonological onset.

Type	Item	Target	Distractor	Type	Item	Target	Distractor
Critical trials	1	Pills	Pillows	Filler trial manipulation	1	Markers	Marbles
	2	Turtles	Turkeys		2	Ladders	Laptops
	3	Sandals	Sandwiches		3	Bows	Boas
	4	Papers	Paints		4	Lemons	Lettuce
	5	Rats	Rabbits		5	Flags	Flashlights
	6	Matches	Maps		6	Beets	Beans
	7	Cards	Cars		7	Berries	Bears
	8	Watermelons	Waffles		8	Canes	Capes
	9	Mushrooms	Muffins		9	Apples	Cookies
	10	Dolls	Dogs		10	Kites	Footballs
	11	Seals	Seagulls		11	Medals	Trophies
	12	Peas	Pizzas		12	Rings	Tiaras
	13	Bees	Beetles		13	Hotdogs	Pies
	14	Baskets	Bats		14	Pipes	Combs
	15	Socks	Soccer balls		15	Carpets	Lamps
	16	Robes	Roses		16	Racquets	Bicycles
Filler trial (other gender – empty set)	1	Ladybugs	Leaves	Filler trial (other gender – subset)	1	Tables	Couches
	2	Bananas	Carrots		2	Cats	Fishes
	3	Gloves	Skates		3	Compasses	Rulers
	4	Bells	Whistles		4	Hammers	Screwdrivers
	5	Purses	Wallets		5	Cameras	Phones
	6	Candy	Cakes		6	Toothpaste	Toothbrushes
	7	Bottles	Cans		7	Knives	Spoons
	8	Plates	Cups		8	Arrows	Bows
	9	Candles	Light bulbs		9	Shoes	Balls
	10	Coins	Keys		10	Suitcases	Back bags
	11	Peppers	Pickles		11	Ties	Cufflinks
	12	Pans	Ladles		12	Sunglasses	Umbrellas

13	Beakers	Books	13	Toasters	Blenders
14	Helmets	Jerseys	14	Speakers	Headphones
15	Belts	Watches	15	CDs	Ipods
16	Drums	Guitars	16	Envelopes	Notebooks

Appendix C. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cogpsych.2018.01.004.

References

Adler, R., Novick, J., & Huang, Y. (2018). Context, conflict, and the time course of interpreting irony. *Paper presented at the 31st annual CUNY conference on Human Sentence Processing*. Davis, California.

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. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*, 247–264.

Arnold, J. E., & Watson, D. G. (2015). Synthesising meaning and processing approaches to prosody: Performance matters. *Language, Cognition, and Neuroscience, 30*, 88–102.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., & Singmann, H. (2015). Lme4: Linear mixed-effects models using Eigen and S4, 2014. *R Package Version, 1*(4).

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

Bergen, L., & Grodner, D. J. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 1450–1460.

Bonnefon, J. F., Feeney, A., & Villejoubert, G. (2009). When some is actually all: Scalar inferences in face-threatening contexts. *Cognition, 112*, 249–258.

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

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

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

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.

Chemla, E., & Spector, B. (2011). Experimental evidence for embedded scalar implicatures. *Journal of Semantics, 28*(3), 359–400.

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

Chierchia, G., Fox, D., Spector, B. (2012). Scalar implicature as a grammatical phenomenon. In: K. von Heusinger, C. Maienborn, & P. Portner (Eds.), Semantics: An international handbook of natural language meaning, Vol. 3, chap. 87, (pp. 2297–2331). Berlin: Mouton de Gruyter.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*, 181–204.

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.

Degen, J., & Tanenhaus, M. K. (2015). Processing scalar implicature: A constraint-based approach. *Cognitive Science, 39*, 667–710.

Degen, J., & Tanenhaus, M. K. (2016). Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive Science, 40*, 172–201.

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*, 20120394.

Dikker, S., & Pylkkanen, L. (2011). Before the N400: Effects of lexical–semantic violations in visual cortex. *Brain and Language, 118*, 23–28.

Dikker, S., & Pylkkänen, L. (2013). Predicting language: MEG evidence for lexical preactivation. *Brain and Language, 127*, 55–64.

Fine, A. B., & Florian Jaeger, T. (2013). Evidence for implicit learning in syntactic comprehension. *Cognitive Science, 37*, 578–591.

Fowler, C. A., & Housum, J. (1987). Talkers' signaling of “new” and “old” words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language, 26*, 489–504.

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

Geurts, B. (2009). Scalar implicature and local pragmatics. *Mind & Language, 24*(1), 51–79.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences, 110*, 8051–8056.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences, 20*, 818–829.

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

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

Hartshorne, J. K., Azar, S. Y. L., Snedeker, J., & Kim, A. (2015). The neural computation of scalar implicature. *Language, Cognition, and Neuroscience, 30*, 620–634.

Hartshorne, J. K., and Snedeker, J. (2014). The speed of inference: Evidence against rapid use of context in calculation of scalar implicature. Unpublished manuscript.

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. (1989). *A natural history of negation.* Chicago, IL: University of Chicago Press.

Huang, Y. & Arnold, J. (in press). Talking about SOME and ALL: What determines the usage of quantity-denoting expressions? To appear in Discourse Processes.

Huang, Y., & Gordon, P. (2011). Distinguishing the time-course of lexical and discourse processes through context, co-reference, and quantified expressions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 966–978.

Huang, Y. T., Newman, R. S., Catalano, A., & Goupell, M. J. (2017). Using prosody to infer discourse prominence in cochlear-implant users and normal-hearing listeners. *Cognition, 166*, 184–200.

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

Huang, Y., & Snedeker, J. (2011). ‘Logic & Conversation’ revisited: Evidence for a division between semantic and pragmatic content in real time language comprehension. *Language and Cognitive Processes, 26*, 1161–1172.

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 824–843.
Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49*, 133–156.
Kim, A., & Lai, V. (2012). Rapid interactions between lexical semantic and word form analysis during word recognition in context: Evidence from ERPs. *Journal of Cognitive Neuroscience, 24*, 1104–1112.
Kowalski, A., & Huang, Y. (2017). Predicting and priming thematic roles: Flexible use of verbal and structural cues during relative clause comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*, 1341–1351.
Ladusaw, W. A. (1994). Thetic and categorical, stage and individual, weak and strong. In M. Harvey, & L. Santelmann (Eds.). *Proceedings from semantics and linguistic theory IV*. Cornell University, Department of Modern Languages and Linguistics.
Lam, T. Q., & Watson, D. G. (2010). Repetition is easy: Why repeated referents have reduced prominence. *Memory and Cognition, 38*, 1137–1146.
Levinson, S. (2000). *Presumptive meanings.* Cambridge, MA: MIT Press.
Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences, 106*, 21086–21090.
Lewis, S. (2013). *Pragmatic enrichment in language processing and development.* Unpublished doctoral dissertationUniversity of Maryland College Park.
Mani, N., & Plunkett, K. (2010). In the infant's mind's ear: Evidence for implicit naming in 18-month-olds. *Psychological Science, 21*, 908–913.
Mani, N., & Plunkett, K. (2011). Phonological priming and cohort effects in toddlers. *Cognition, 121*, 196–206.
Matin, E., Shao, K. C., & Boff, K. R. (1993). Saccadic overhead: Information processing time with and without saccades. *Perception and Psychophysics, 53*, 372–380.
Meyer, A. S., Belke, E., Telling, A., & Humphreys, G. W. (2007). Early activation of object names in visual search. *Psychonomic Bulletin and Review, 14*, 710–716.
Nieuwland, M., Ditman, T., & Kuperberg, G. (2010). On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities. *Journal of Memory and Language, 63*, 324–346.
Noveck, I. A., & Posada, A. (2003). Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language, 85*, 203–210.
Panizza, D., Chierchia, G., Huang, Y., & Snedeker, J. (2009). Relevance of polarity for the online interpretation of scalar terms. *Semantics and Linguistic Theory, 19*, 360–378.
Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36*, 329–347.
Politzer-Ahles, S., & Fiorentino, R. (2013). The realization of scalar inferences: Context sensitivity without processing cost. *PLoS ONE, 8*, e63943.
Postal, P. (1964). Limitations of phrase structure description. In J. K.a. J. Fodor (Ed.). *Readings in the Philosophy of Language*. Englewood Cliffs, NJ: Prentice-Hall.
Rees, A. & Bott, L. (in press). The role of alternative salience in the derivation of scalar implicatures. To appear in *Cognition*.
Rips, L. J. (1975). Quantification and semantic memory. *Cognitive Psychology, 7*, 307–340.
Rubio-Fernández, P. (2016). How redundant are redundant color adjectives? An efficiency-based analysis of color overspecification. *Frontiers in Psychology, 7*, 153.
Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research, 32*, 3–23.
Skordos, D., & Papafragou, A. (2016). Children's derivation of scalar implicatures: Alternatives and relevance. *Cognition, 153*, 6–18.
Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*, 1632–1634.
Tomlinson, J. M., Bailey, T. M., & Bott, L. (2013). Possibly all of that and then some: Scalar implicatures are understood in two steps. *Journal of Memory and Language, 69*, 18–35.
Watson, D. G. (2010). The many roads to prominence: Understanding emphasis in conversation. *Psychology of Learning and Motivation, 52*, 163–183.
Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology, 58*, 250–271.
Zelinsky, G., & Murphy, G. (2000). Synchronizing visual and language processing: An effect of object name length on eye movements. *Psychological Science, 11*, 125–131.