Another look at the online processing of scalar inferences: an investigation of conflicting findings from visual-world eye-tracking studies

Chao Sun & Richard Breheny

Routledge
Taylor & Francis Group

REGULAR ARTICLE

🔓 OPEN ACCESS Check for updates

Another look at the online processing of scalar inferences: an investigation of conflicting findings from visual-world eye-tracking studies

Chao Sun[a,b] and Richard Breheny[b]

aDepartment of German Language and Linguistics, Humboldt-Universität zu Berlin, Berlin, Germany; bDivision of Psychology and Language Sciences, University College London, London, UK

ABSTRACT
Previous psycholinguistic studies that compared the time course of interpretation for *pragmatic some* and *literal all* have returned mixed results. In particular, a delayed *pragmatic some* has been found in some studies but not in others. We explain these conflicting findings in terms of factors which are independent of incremental semantic/pragmatic interpretation. Two offline experiments provide evidence of the effect of these factors. Three visual-world studies showed that they influence participants' eye movements in online comprehension. We introduce a new measure for investigating the time course of scalar inference. This new measure allows us to reason about the time course question based on the difference in verification procedures between numbers and quantifiers. Results from our visual-world studies suggest that deriving the pragmatic interpretation is not delayed relative to the semantic interpretation.

Introduction

What speakers mean is often underspecified in what they say. Listeners frequently make extra-linguistic inferences to enrich the message during language comprehension. Here we focus on the pragmatic enrichment associated with the quantifier *some*.

(1) I ate some of the cookies.

The semantic interpretation of *some* can be paraphrased as "some and possibly all". Thus, the literal meaning of (1) does not rule out the possibility that the speaker ate all of the cookies. However, in many contexts, the speaker who uttered (1) implies that she did not eat all of the cookies. *Some* then receives a pragmatic interpretation "some but not all" which excludes the situation that is compatible with *all*. In what follows, we will sometimes use the terms *semantic some* and *pragmatic some* to refer to semantic and pragmatic interpretations of *some*.

The pragmatic enrichment from *some* to "some but not all" is a representative example of so-called scalar inference (SI), and although there is controversy about the status of SIs, many believe that it involves a pragmatic enrichment of literal meaning.[1] To the extent that SI is pragmatic, researchers, following on from Grice (1967, 1989) explain SI in terms of an inference about the speaker who

asserts the literal proposition. Thus according to Grice's conceptual framework, establishing the literal meaning takes conceptual priority over deriving conversational implicatures. That is to say, first the literal meaning is derived based on the sentence structure and semantic rules, then pragmatic enrichments are calculated based on the literal meaning and conversational maxims. When implementing Grice's conceptual framework to actual language processing, it raises the question of whether the semantic processing also has temporal priority over pragmatic processing.

Two different predictions have been discussed in the processing literature. One is the *slow pragmatic / literal first* view, which proposes that in on-line processing the semantic interpretation is accessed prior to deriving pragmatic inferences (Huang & Snedeker, 2009). The other one is the *fast pragmatic* view, which argues that the conceptual priority of establishing the semantic interpretation is not necessarily mapped onto a temporal priority, i.e. integrating pragmatic inferences is not inevitably delayed relative to accessing the semantic interpretation (Breheny, Ferguson, & Katsos, 2013; Degen & Tanenhaus, 2015, 2016; Grodner, Klein, Carbary, & Tanenhaus, 2010; Politzer-Ahles & Fiorentino, 2013).

In order to test these two different on-line predictions, the time course of scalar inference has been studied as a

---

test case. In particular, using the visual-world eye-tracking paradigm, psycholinguistic studies compared the timing of integrating *pragmatic some* with the timing of accessing the semantic interpretation of *all*. Some of these studies found a delay in integrating the pragmatic inference relative to the semantic interpretation (Huang & Snedeker, 2009, 2011, 2018), whereas others reported a rapid integration (Breheny et al., 2013; Grodner et al., 2010).

One goal of this paper is to provide an explanation for previous inconsistent findings on the time course of interpretation. A further goal is to re-examine the time course of scalar inference using a novel visual-world eye-tracking paradigm. The outline of this paper is as follows. We will first review previous visual-world studies on the time course of scalar inferences. This will be followed by a discussion of issues that make the interpretation of previous eye movement data problematic. We suggest that processing differences attested between *all* and *pragmatic some* can be due to two factors which relate specific set size and determiner: prior associations between relative set sizes and quantifiers and Maximise Presupposition. This account will be tested empirically in two rating experiments (Experiments 1a and 1b) and three visual-world studies (Experiments 2a, 2b and 3). Experiments 1a and 1b provide an independent test of the effect of these factors on participants' expectations about visual referents. Experiments 2a,b and 3 test whether relative set size affects comprehenders' eye movement during the on-line processing of quantificational expressions. The stimuli for experiments 2a,b and 3 also introduce a distinctive visual area, the residue set. It is visual attention to this area that allows us to determine the time course of meaning composition independently of low-level associations.

Background

In a series of innovative visual world studies, Huang and Snedeker (2009, 2011) reported data that supports the *slow pragmatic* view. In their studies, participants were presented with visual displays depicting a girl with all of one kind of item (e.g. three soccer balls), a girl with some but not all of another kind of item (e.g. two of four socks), and some distractors, while they listened to a sentence such as "Point to the girl that has *some/two* of the socks" or "Point to the girl that has *all/three* of the soccer balls". The question is, will participants shift their looks towards the correct referent upon hearing the quantifier or number word? Crucially, upon hearing *some*, the slow pragmatic view predicts that there should be looks to both girls as the semantic interpretation of *some*, which is compatible with both girls, is

accessed prior to implicature calculation. By contrast, the fast pragmatic view predicts that, if contextual support is sufficient, participants should rapidly shift their looks to the girl with some but not all of the socks as the immediate integration of *pragmatic some* would exclude the girl that is compatible with *all*. Huang and Snedeker found that participants initially looked equally at both girls, which led to a delay in identifying the referent in *some* compared to conditions where no pragmatic inference was involved. In particular, visual preferences to the correct referent emerged immediately after the onset of *all, two* and *three*, but only emerged approximately 800ms after the onset of *some*. They interpreted the delay in *some* as evidence in support of the slow pragmatic view that the pragmatic interpretation is preceded by accessing the semantic interpretation in the early stage.

However, other similarly constructed visual-world studies reported data that support the fast pragmatic view (Breheny et al., 2013; Grodner et al., 2010). For instance, Grodner and colleagues showed no delay in referential disambiguation based on *pragmatic some* relative to *all*. The rapid integration of *pragmatic some* found in these studies is also consistent with previous research showing that effects of contextual inference are not necessarily delayed relative to the effects of semantic composition, even where the contextual inference is based on Gricean reasoning (Altmann & Steedman, 1988; Sedivy, 2003; Sedivy, Tanenhaus, Chambers, & Carlson, 1999; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995).

In summary, data from visual-world eye-tracking studies are inconclusive about whether on-line processing of scalar inference is delayed. It is also unclear what factors lead to these conflicting findings. In the next section, we will first review studies that attempt to account for the discrepant results. We will then discuss several factors that have not been explored yet but may make the interpretation of the previous eye-tracking data problematic.

Explaining the conflicting findings

Comparing studies that found little or no delay in integrating *pragmatic some* to those that found a delay, one key difference is that the referent in the latter case was also described using number words. For example, in the course of a session for a participant, the target for "the girl that has *some* of the socks" would also be described as, "the girl that has *two* of the socks". Thus, recent studies that aim to explain previous conflicting findings have been focused on the role of these number items.

Two different hypotheses have been put forward. The verbal-encoding hypothesis, proposed by Huang and Snedeker (2018), postulates that if there is only one way of referring to the referent, participants may implicitly pre-label the visual referent and such labelling can speed up referent identification in visual-world studies. Consider visual-world studies without numbers, where the referent is only described using quantifier phrases (e.g. Click on the girl who has *all/some* of the balloons). In this situation, participants could easily predict how referents would be described in terms of quantifiers and pre-code the visual display prior to the instruction onset, such as labelling the total sets with *all* and the subsets with *some*. Verbal encodings of this sort would facilitate referent identification for both *all* and *some*, so that upon hearing the quantifier, participants could immediately fixate on the referent of which the pre-coded label matches the audio description. However, this top-down strategy will bypass bottom-up analysis including establishing the semantic meaning in both conditions and deriving the *not all* inference in *some* condition. Thus, Huang and Snedeker (2018) suggest that no difference between *some* and *all* found in studies without numbers could be the result of verbal-encoding rather than the rapid integration of scalar inference.

By contrast, consider visual-world studies with numbers, where the referent is described by both quantificational and numerical scales (e.g. girl that has *all/three* of the socks). Huang and Snedeker argue that verbal-encoding should be discouraged because it is inefficient to pre-code the referent with a single type of description.

Huang and Snedeker (2018) provide empirical evidence supporting the verbal-encoding hypothesis. Using the same paradigm as Huang and Snedeker (2009), they conducted a between-subjects study manipulating the presence of number instructions. They found that referent identification for *some* is delayed relative to *all* only when numbers are present but not when they are absent. This finding is in line with the prediction of the verbal-encoding hypothesis. Therefore, according to Huang and Snedeker (2018), the processing of pragmatic *some* is still delayed, when there is no delay associated with pragmatic *some*, verbal-encoding is at play.

An alternative hypothesis for explaining the disparity (i.e. experimental data for and against the delayed processing of *pragmatic some*) could be termed the naturalness hypothesis (Degen & Tanenhaus, 2015, 2016; Grodner et al., 2010). The proposal is that the presence of number words undermines the naturalness of *some* used with a small (subitisable) set, and the unnatural use of the quantifier delays the processing of scalar inference. Grodner et al. (2010) demonstrated in an off-line naturalness rating task that when both *some* and *all* were used to describe small sets (i.e. two or three items), the presence of numbers reduced the naturalness of *some*. Thus, they speculated that processing differences observed in Huang and Snedeker (2009) could be explained as a reflection of naturalness differences between *some* and *all* when numbers are present.

Degen and Tanenhaus (2015) provided further evidence suggesting that, in off-line judgments, the naturalness of *some* varies with the presence or absence of numbers. In addition, the effect of numbers differs for different set sizes. In a series of rating studies, they found that the presence of numbers lowered the naturalness of *some* when used for small sets (1–3 gumballs), but not to the same extent when used for big sets (e.g. 4–5 gumballs). In Huang and Snedeker (2009), *some* was always used with a set of two items. Thus, Degen and Tanenhaus speculated along the same lines as in Grodner et al. (2010) that the delay in *pragmatic some* might be partly caused by the unnatural use of the quantifier *some*.

The studies in Degen and Tanenhaus (2016) provide a test for predictions made by both the verbal-encoding hypothesis and the naturalness hypothesis. Two eye-tracking experiments investigated the time course of integrating *pragmatic some* while manipulating (i) set size and (ii) the availability of number words. Participants were presented with a gumball machine of which the lower chamber contained a total set of gumballs of one colour (e.g. four orange gumballs) and a subset of gumballs of another colour (e.g. two of eight blue gumballs). While viewing the display, they listened to a sentence describing the lower chamber such as "You got *all/four* of the orange gumballs" or "You got *some/two* of the blue gumballs". Participants' task was to click on the set of gumballs described in the sentence. Set size was manipulated by using quantifiers, *some* and *all*, equally often to refer to both big sets (i.e. 4 or 5 gumballs) and small sets (i.e. 2 or 3 gumballs). For instance, a display could depict that the lower chamber contains a total set of two orange gumballs and a subset of four of the eight blue gumballs. This display could be paired with sentences such as "You got *all/two* of the orange gumballs" or "You got *some/four* of the blue gumballs". The availability of number words was manipulated by excluding numbers in their first experiment and then including numbers in the second experiment.

The verbal-encoding hypothesis predicts that regardless of set size, there should be no difference in looking patterns between *some* and *all*[2] in the number *absent* study, whereas in the number *present* study visual preference to the target set for *some* should be delayed relative

to *all*. The naturalness hypothesis makes on-line predictions based on the off-line naturalness judgements. Consider the number *absent* study. Previous research has shown that when numbers are not available, the naturalness of *some* and *all* does not differ (Degen & Tanenhaus, 2011; van Tiel, 2014). Thus, in terms of online measures, the naturalness hypothesis predicts that in the number *absent* study looking patterns between *some* and *all* should not differ either. Then consider the number *present* study. By hypothesis, the presence of numbers reduces the naturalness of quantifier use most severely when *some* is used with small sets. Thus, in terms of eye movements, the naturalness hypothesis predicts that, in the number *present* study, visual preference to the target set for *some* should be delayed relative to *all* when the target set is small, but not when it is big.

In their *number absent* study, Degen and Tanenhaus found that there was no difference in looking patterns between *some* and *all*. These results are predicted by the naturalness hypothesis; however, they are also consistent with the verbal-encoding hypothesis which predicts that participants label the total set with *all* and the subset with *some* and consequently identify the visual target rapidly in both conditions. Degen and Tanenhaus argue that their findings are less likely to be explained by the verbal-encoding hypothesis due to the use of garden path trials. In those trials, participants were presented either with a semantically false statement or a pragmatically infelicitous statement. An example of the former could be a statement "You got all of the blue gumballs" paired with a display showing that the lower chamber contains only a subset of the blue gumballs. An example of the latter could be a statement "You got some of the blue gumballs" paired with a display showing that the lower chamber contains a total set of the blue gumballs. In these trials, participants were instructed to click on a central button on the gumball machine if they judged the statement to be false. For a quarter of the *some* trials, *some* was used with a total set and the same for *all* used with a subset. Thus, Degen and Tanenhaus argue that, the potential for pre-coding (i.e. labelling the total set with *all* and the subset with *some*) is lower in their number *absent* study, compared with other previous studies like Grodner et al. (2010) and Huang and Snedeker (2018, Exp. 2) where pre-coding would be more effective.[3] Taken together, Degen & Tanenhaus's number *absent* study adds support to the fast pragmatic account found in Grodner et al. (2010), without completely ruling out the verbal-encoding hypothesis.

In their number *present* study, Degen and Tanenhaus found that visual preference to the target set for *some* was delayed relative to *all* when the target set was big,

but not when it was small. This interaction between set size and quantifier use is not predicted by the slow pragmatic/verbal-encoding hypothesis. Interestingly, it is not predicted by the naturalness hypothesis either. In fact, this finding is the opposite to what the naturalness hypothesis predicted. When *all* and *some* were used with a big set, the observed delay cannot be attributed to the effect of numbers. This is because the presence of numbers was expected to reduce the naturalness of *all* and *some* approximately to the same extent. Conversely, when *all* and *some* were used with a small set, the no-difference looking pattern was unexpected because the naturalness of *some* used with small sets was affected the most by the presence of numbers. Although these results cannot be straightforwardly explained by the two hypotheses on the market, the interaction between set size and quantifier use still poses problems for the slow pragmatic view. The slow pragmatic view suggests that implicature calculation happens after the formation of the semantic interpretation. Therefore, it predicts that *pragmatic some* is delayed relative to *all* regardless of the target set size. However, the interaction found in Degen and Tanenhaus's number *present* study clearly showed that this temporary delay in *pragmatic some* was only observed when the target set was big.

In summary, two hypotheses, with special focus on the role of numbers, have been proposed to account for the conflicting findings in visual-world studies on the processing of scalar inference. The slow-pragmatic/verbal-encoding hypothesis explains the rapid integration of *pragmatic some* as the result of labelling the target with quantifiers beforehand which is encouraged by lack of number items. Whereas the naturalness hypothesis attributes the delay associated with *pragmatic some* to the unnatural use of *some* with small sets caused by the availability of number alternatives. However, data from recent eye-tracking studies are not in line with these two hypotheses about the discrepant findings. In what follows, we propose another account of the delay/no-delay findings, which could also account for the puzzling interaction between set size and quantifier use.

In the visual-world studies reviewed above, participants' eye movements were considered to primarily reflect the incremental processing of the linguistic stimuli. However, many studies have shown that anticipatory eye movements are influenced by factors that go beyond the linguistic lexical information, such as world knowledge (e.g. Altmann & Kamide, 2007) and prosody (e.g. Kurumada, Brown, Bibyk, Pontillo, & Tanenhaus, 2014). Thus, participants' eye movements collected from studies on the time course of scalar inference might reflect other factors in addition to compositional

language processing. One possible factor concerns a low-level association between the quantifier *all* and the larger set in the display. That is to say, given an instruction with *all*, participants would expect *all* to be used with the larger set as there was always a contrast between a larger set and a smaller set in visual displays in previous studies. Such an association could influence participants' anticipatory eye movements in a way that it facilitates target identification if *all* is used with the larger set but interferes with it if *all* is used with the smaller set. This effect could be at least as early as, if not prior to, the effect of the composition of the meanings (whether pragmatically enriched or not) of the linguistic expressions used.

If this low-level association is indeed in play, it could partly account for the interaction found in Degen & Tanenhaus's number *present* study. When the target was the smaller set, the association between *all* and the larger set would drive looks away from the target, consequently, target bias in *all* trials would be disadvantaged compared to *some* trials. This could lead to an advantage in target identification for *some* over *all*. Whereas when the target was the larger set, assuming the association with the larger set was stronger for *all* than for *some*, then target identification in *all* trials would be facilitated compared to *some* trials. This would explain the early target bias formation for big *all* trials compared to big *some* trials.

Another factor that might affect eye movements in previous visual-world studies is related to Maximise Presupposition (Heim, 1991; Sauerland, 2008). The maximise presupposition principle blocks the use of an utterance when an alternative utterance has a more informative presupposition.[4] In Huang and Snedeker (2009) and Grodner et al. (2010), "the girl that has all of … " was paired with a display containing one girl that has two objects and another girl that has three or more.[5] Given that *all* has a weaker existential presupposition, it is less compatible with the girl that has two objects, for which the presupposition of *both* would be the better choice. If Maximise Presupposition is indeed applied, then upon hearing *all*, this conflict would drive looks away from the girl with a set of two objects, independent of whether or not this girl is the referent of *all*. Therefore, when *all* was used and one set of items in the display has cardinality 2, target identification in *all* trials would be delayed if *all* is used with this set of two objects, and it would be facilitated if *all* is used for the other set that has cardinality larger than 2.

In Experiments 1a and 1b below, we seek independent motivation for our suppositions concerning previous studies, that low-level associations between quantifiers and relative set size and Maximise Presupposition affect preferences independently of incremental verification processes exploiting determiner meaning. In Experiments 2a and 2b, we test on-line predictions based on our suppositions in visual-world experiments.

Regardless of factors that may contribute to the conflicting results, the fundamental question remains: is integrating scalar inference delayed relative to semantic interpretation? Here we discuss two issues that have been ignored when interpreting previous eye-tracking data. First, a no-difference result in looking patterns between *all* and *numbers* is puzzling. Consider Huang and Snedeker (2009). The process of identifying the referent of the description "the girl that has *all/three* of the … " involves verifying the relative clause ("x has *all/three* … ") against the sets of objects associated with the girls in the display. For *all* trials, in order to establish that the girl with the three soccer balls is the *all* referent, the whole visual display needs to be checked to ensure that no other characters (e.g. either of the boys) obtained any soccer balls. However, for *number* trials, it is sufficient to only inspect the cardinality of the set that each girl has. Thus, given the difference in the region of inspection required to anticipate the referent, we would expect that target gaze bias should build faster in *numbers* than in *all*. Yet, Huang and Snedeker (2009) found no difference in looks to the target between *all* and *three* after the onset of the determiner. Given a low-level association of the kind discussed above and potential Maximise Presupposition effects, it is plausible that in Huang and Snedeker's study the expected delay based on the different verification processes between *all* and *numbers* is compensated for by the factors which drive early gaze bias to larger targets in *all* trials.

Another issue arising from previous visual-world studies is that they were designed in a way that the fast pragmatic view would predict an absence of effect. Consider Grodner et al. (2010). The fast pragmatic view would predict no difference in looks to the target between *some* and *all*. Grodner and colleagues indeed found the null result that was consistent with the prediction, but they had to demonstrate in post-analyses that it was reasonable to accept the null hypothesis.

These two issues will be addressed in our visual-world paradigm. We offer a novel dependent measure which allows us to measure gaze formation differences that we should expect if participants use the compositional meaning of *all* as against *numbers*. In addition, this novel dependent measure allows the fast pragmatic view to formulate an alternative hypothesis regarding the time course question. More detail about our visual-world paradigm will be given in Experiments 2a, b and 3. Tables 1–3 summarise and compare different predictions regarding target preference and visual search to the residue set.

Table 1. Summary of predictions regarding the effect of *Determiner* on target preference in Experiment 2a, 2b and 3, collapsing over target set size.

	Experiment 2a, 2b and Experiment 3	
Account	Intercept (overall target preference)	Linear slope (rate of increase)
Slow pragmatic	all > some	all > some
	number = all, number > some	number = all, number > some
Fast pragmatic	all = some	all = some
	number > all, number > some	number > all, number > some

Notes: > indicates that the target preference is greater in one condition than the other, = indicates predicted no difference. Note that *Slow Pragmatic* reflects assumptions in Huang and Snedeker (2009), among others, that set size has no effect and that target bias in *number* trials and *all* trials should be equivalent.

Table 2. A comparison between predictions regarding the effect of *Target set size* on target preference.

Comparison	*Slow pragmatic*	Exp. 2a and Exp. 3 Prediction	Exp. 2b Prediction
big-set *all* vs. small-set *all*	=	>	>
big-set *some* vs. small-set *some*	=	>	=
big-set *all* vs. big-set *some*	>	>	>
small-set *all* vs. small-set *some*	>	<	<

Notes: > and < indicates the presence and direction of the predicted effect. > indicates that the target preference is greater in one condition than the other, and < indicates the target preference is smaller in one condition than the other. = indicates predicted no difference. Note that *Slow Pragmatic* reflects assumptions in Huang and Snedeker (2009), among others, that set size has no effect and that target bias in *number* trials and *all* trials should be equivalent. Our predictions are based on the assumption that Maximise Presupposition and a low-level association between determiner and set size influence bias, plus that the enriched interpretation of *some* is as fast as *all*.

Experiments 1a and 1b

Experiments 1a and 1b were designed to independently test our conjecture that, given the relative set sizes used in previous studies, participants form a preference for a target independently of critical linguistic information, based on low-level associations for quantificational determiners and/or Maximise Presupposition. Our test here uses off-line rating tasks. We constructed simple

Table 3. Summary of predictions regarding the effect of *Determiner* on visual search to the residue set in Experiment 2a, 2b and 3.

	Experiment 2a, 2b and Experiment 3		
Account	Intercept	Linear slope	Quadratic slope
Slow pragmatic	number < all, number np some, some < all	number < all, number = some	number: no, some: no, all: yes
Fast pragmatic	number < all, number < some, some = all	number < all, number < some	number: no, some: yes, all: yes

Notes: < indicates the bias to the residue set is smaller in one condition than the other, = indicates predicted no difference, and np indicates no predicted effect. Note that for Quadratic slope, the prediction is about the presence or absence of a U-shaped curve, not a comparison between conditions.

sentences containing *some* and *all*, e.g. "The girl has some of her sister's flowers" or "The boy has all of his cousin's candies". These sentences were paired with a slider-rating scale with one image located on each end (Figure 1). Both images depict a character with a set of objects, but neither are clear on whether the character possesses a total set or a subset. Crucially, two sets of objects differ in the number of items contained, so that there is always a contrast between the larger set and the smaller set. In both experiments, participants were asked to indicate which image fits better with the sentence by moving the slider towards the chosen image. There was no clear solution of the task. However, a low-level association between the determiner *all* and the larger set would predict that for an *all* sentence, participants move the slider towards the image depicting the character with a larger set. Moreover, if the association between *all* and the larger set is stronger than the one between *some* and the larger set, it would predict that participants move the slider further towards the image with the larger set for an *all* sentence than for a *some* sentence.

Experiments 1a and 1b differ in the relative proportions of numbers used. In Experiment 1a, characters have either a set of two objects or three. These quantities were chosen specifically because they were used in Hang & Snedeker's studies, with *all* targets always holding a set of three, while *some* targets having two. In Experiment 1b, characters have three objects or four. These quantities allow us to determine if an association between *all* and set size has an effect independently of Maximise Presupposition.

Experiment 1a

Participants
Sixty-nine participants were recruited via Amazon Mechanical Turk. They were asked to indicate their native language and only participants with English as a native language were included in the analyses.

Materials and procedure
Figure 1 (top) is an example item. We constructed two target sentences: "The girl has some of her sister's flowers" and "The boy has all of his cousin's candies". For target items, two pictures differed in set sizes such that the larger set contained three items and the smaller set contained two items. Participants were instructed to use the slider to indicate their intuition about which image fits better with the sentence; such that, if one image is a lot better than the other, they move the slider right over toward that image, and if one image is only a little bit better, then they may move the slider a
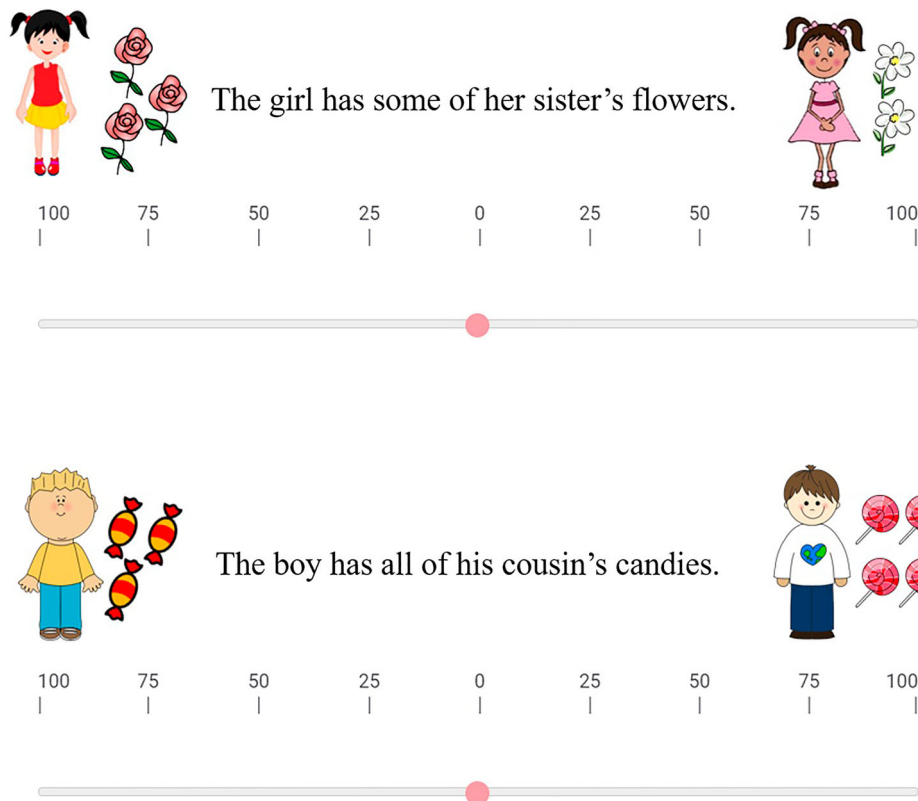
Figure 1. Examples for experimental items used in Experiment 1a (top) and in Experiment 1b (bottom).

little bit toward that image. The slider's starting point was placed at the centre of the scale. 4 control items were constructed, of which two were clearly unambiguous on which image they described, and two were clearly ambiguous. Example items are provided in Appendix A.

Each participant judged 2 target items and 4 control items. 8 lists with the pseudo-randomised order of trials were created. This was to ensure that overall for each target sentence, the larger set and the smaller set were equally often to appear at two ends of the slider scale. In addition, the objects (e.g. rose, daisy) that constituted the larger set and the smaller set were counterbalanced within each quantifier. Participants were randomly assigned to one of the eight lists.

Results and discussion

Responses from 11 non-native English speakers and 19 participants who made mistakes on control items were excluded. 39 participants were analysed. For each trial, participants' slider-rating was mapped on to a 0 (smaller set) −200 (larger set) continuous scale, where 100 corresponds to the slider starting anchor point. A rating of 100 indicates no preference for one image over the other. To test whether there were any low-level associations between quantificational determiners and relative set sizes, for each target sentence, we

compared mean ratings with 100 using one-sample t-tests. We found that the mean rating for "the girl has all of her sister's flowers" was significantly higher than 100 ($M = 131.69$, t (38) = 6.69, $p < .001$). Interestingly, the mean rating for "the boy has some of his cousin's candies" was also significantly higher than 100 ($M = 120.85$, t (38) = 3.763, $p = .001$). A paired-sample t-test was conducted to compare participants' ratings for *all* and *some* sentences. We found that the mean rating for the *all* sentence was significantly higher than that for the *some* sentence (t (38) = 2.35, $p = .024$).

Thus, for both target sentences, participants preferred the image depicting the girl with a larger set of three. However, the preference for *all* to be used with the larger set was stronger than the preference for *some* to be used with the larger set. Note that a statistically significant preference between *some* and the larger set was not predicted beforehand. A plausible explanation for this finding is that participants had a dis-preference for *some* to be used with a set of two objects compared to a set of three. This explanation reflects preferences reported in Degen and Tanenhaus (2015).

Regarding the preference in the *all* case, given the discussion above, this could be due to the association between determiner and larger set, or a dis-preference for *all* to be used with a set of two, or both factors.

Experiment 1b had two aims. The first was to determine the robustness of the low-level association independently of the Maximise Presupposition principle. We increased the cardinality of the object sets so that the larger set was a set of four and the smaller set was a set of three. According to Maximise Presupposition, the *all* sentence was felicitous to describe both images on the slider-rating scale. If there is a low-level association between *all* and the larger set, we would expect that participants move the slider towards the image depicting the girl with four objects. The second aim was to test our explanation for the finding of the *some* sentence in Experiment 1a. If this finding is due to the dis-preference for two objects rather than any association between *some* and the larger set, we would expect that participants show no clear preference for either the larger or the smaller set in Experiment 1b.

Experiment 1b

Participants
Fifty-two participants were recruited via Amazon Mechanical Turk. They were asked to indicate their native language and only participants with English as a native language were included in the analyses.

Materials and procedure
Figure 1(bottom) is an example item. The materials were similar to Experiment 1a with one key difference. In the display, the larger set was a set of four and the smaller set was a set of three. The procedure was identical to Experiment 1a.

Results and discussion
Responses from 2 non-native English speakers and 12 participants who made mistakes on control items were excluded. 38 participants were analysed. Participants' ratings were mapped on to a 0 (smaller set) −200 (larger set) continuous scale. As we did in Experiment 1a, for each target sentence, we compared mean ratings with 100 using one-sample t-tests. We found that the mean rating for "the girl has all of her sister's flowers" was significantly higher than 100 ($M = 115.32$, $t(37) = 4.99$, $p < .001$), but the mean rating for "the boy has some of his cousin's candies" was not significantly different from 100 ($M = 97.50$, $p = .56$). The paired-sample t-test showed that the mean rating for the *all* sentence was significantly higher than that for the *some* sentence ($t(37) = 3.29$, $p = .002$).

Thus, for the *all* sentence, participants showed an expected preference for the larger set. Crucially, since there was no violation of the principle Maximise Presupposition, this preference reflects a low-level association between *all* and the larger set. Unlike Experiment 1a, for the *some* sentence, participants showed no clear preference for either set size. This finding argues for the explanation discussed in Experiment 1a that there is no predicted association between *some* and the larger set, though *some* is dis-preferred to be used with a set of two objects.

We did a post hoc analysis to test whether the strength of the association between *all* and the larger set varied between experiments. We constructed a linear regression model predicting ratings based on quantifier (all or some) and experiment (1a or 1b). The analysis revealed a significant effect of quantifier on ratings ($\beta = 12.58$, SE = 5.11, $p = .02$), with greater preference to the larger set in the *all* trials compared to the *some* trials. There was also a significant effect of experiment on ratings ($\beta = 17.74$, SE = 5.11, $p < .001$), with greater preference to the larger set in Experiment 1a compared to Experiment 1b. There was no interaction between quantifier and experiment.

Summary of Experiments 1a and 1b

Previous research on the time course of "some" and "all" has provided evidence suggesting that the relative set sizes in target areas have affected responses independently of the compositional meaning of the determiners involved. We designed Experiments 1a,b to explore this hypothesis further. Our design tasked participants with expressing a preference for one of two potential referents for a definite description when the stimulus provided insufficient information for them to use the linguistic meaning to determine what the referent is. Taken together, the findings provide supporting evidence for our idea that some combination of low-level associations between determiner and set size and Maximise presupposition affects participants' expectations about reference. In fact, we found that there is not only an association between the determiner *all* and the larger set, the association between *all* and the larger set is also stronger than any association between *some* and the larger set. Moreover, our findings also suggest that the use of two objects creates problems for both determiners in off-line judgement tasks. Specifically, the determiner *all* is less felicitous with a set containing two objects for which "both" would be expected to be a relevant alternative, and the determiner *some* shows clear dis-preference to be used with a set of two objects.[6]

Experiment 2

One aim of Experiments 2a,b is to systematically test whether relative set-size influences participants'

anticipatory looks during on-line comprehension, using the visual-world eye-tracking paradigm. To this end, Experiment 2 consists of two parts. In Experiment 2a, the larger sets in the visual display contained three items and the smaller sets contained two items, whereas in Experiment 2b, the larger sets contained four items and the smaller sets contained three items.[7] Experiment 2 further aims to investigate the question of whether integrating scalar inference is delayed relative to semantic interpretation. As in previous studies, we compared the time course of referential disambiguation based on *pragmatic some* with the time course of referential disambiguation based on *all* and numerical determiners. Differing from previous studies, visual displays in Experiment 2 always include a region where the residue of any partitioned sets of objects remains. This is what we call the "residue set". As explained in detail below, to address the time course question we also compared anticipatory looks to the residue set after hearing quantifiers and numerical determiners.

Experiment 2a

In Experiment 2a, on each trial, participants first saw either Figure 2 (left) or Figure 3 (left), showing four identical cartoon characters and a collection of four sets of objects. After hearing an audio description of the display, the objects were distributed among the characters. While looking at either Figure 2 (right) or Figure 3 (right), participants were listening to a sentence such as (2) or (3). Using two different initial displays allowed us to manipulate the target set size for the *all* and the *some* referent. For instance, (2) can be presented with either Figure 2 (right) referring to a character with a set of three objects or Figure 3 (right) referring to a character with a set of two objects. The opposite applied to (3).

(2) Click on the girl that has *all* of the stripy circles.
(3) Click on the girl that has *some* of the stripy squares.

We were interested in two critical time windows: the determiner window and the modifier window. The determiner window starts from the onset of the determiner and ends before the onset of the modifier (e.g. *some of the*), and the modifier window starts from the onset of the modifier and ends before the onset of the disambiguating noun (e.g. *stripy*). If the hypothesised low-level associations and Maximise Presupposition influence participants' visual attention, then for both the *all* and the *some* condition, the target bias should be greater when the quantifier is used with a larger set (i.e. a set of three objects) compared to when it is used with a smaller set (i.e. a set of two objects). Moreover, since the association between *all* and the larger set is stronger, it should lead to a greater target bias in the *all* condition compared to the *some* condition when used with a larger set.

To lower the possibility of verbal-encoding discussed in Huang and Snedeker (2018), we included number items. For example, the target for "the girl that has *all* of the stripy circles" in Figure 2 (right) was also described as, "the girl that has *three* of the stripy circles". Also, for reasons mentioned above, the inclusion of number items can play an important role in determining the time course of compositional sentence interpretation. That is, verification procedures required for "the girl that has all of … " and "the girl that has three of … "
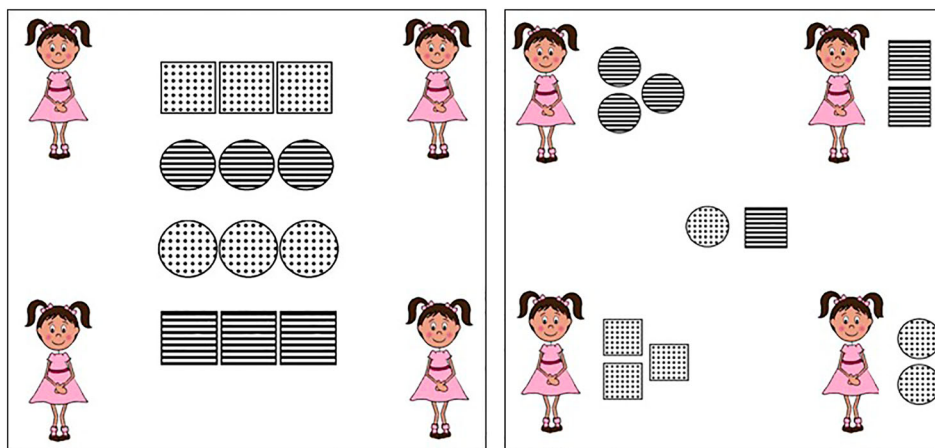

Figure 2. Example displays for Experiment 2a: big *all* / small *some*. The left image is the initial display, and the right image is the critical display. Figure 2 (right) can be paired with "Click on the girl that has all/three of the stripy circles" or "Click on the girl that has some/two of the stripy squares".
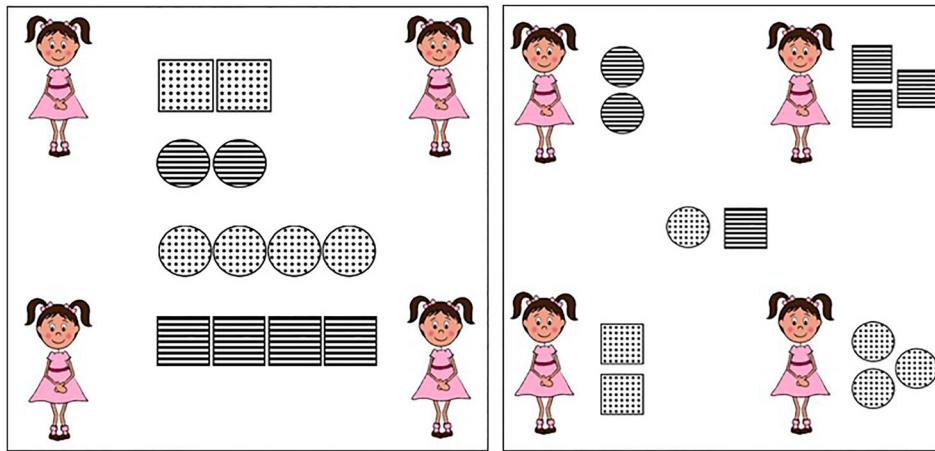
Figure 3. Example displays for Experiment 2a: small *all* / big *some*. The left image is the initial display, and the right image is the critical display. Figure 3 (right) can be paired with "Click on the girl that has all/two of the stripy circles" or "Click on the girl that has some/three of the stripy squares".

should be different. In the current study, we included a residue set in our visual display. As shown in Figure 2 (right) and Figure 3(right), the residue set is located in the centre of the display, consisting of the objects left over after the distribution. Visual search to the residue set should reflect different verification procedures required by *all*, *some* and *numbers*. Specifically, in order to determine whether a character has *all* of a set of objects or whether a character has *some but not all of* a set of objects, participants need to check this residue set. By contrast, to determine whether a character has *two/three* of a set of objects or whether a character has *at least some* of a set of objects, it is sufficient for participants to only check the objects in the character's target region – not the residue region.[8]

Thus the main predictions regarding looks to the residue set relate to comparisons between *all* trials and *number* trials, on the one hand, and between *some* trials and *number* trials on the other hand. Given their linguistic meanings, we predict a greater overall average bias towards the residue set for *all* trials than for numbers. Moreover, we should expect that participants would engage in search behaviour that involves first looking at target areas to see what kinds of objects are there (e.g. stripy circles and dotted squares) and then inspecting the residue set to see which of these kinds of objects is the partitioned set and which is not. I.e. we should expect an initial decrease followed by an increase in looks to the residue set in the *all* condition compared to the *number* condition. The fast pragmatic view, which posits the rapid integration of *pragmatic some*, predicts the same search pattern for *some* as for *all* and the same difference in visual biases between the *some* and the *number* condition. Note that in this way the fast pragmatic view formulates a positive

hypothesis for the time course of *pragmatic some* in addition to predicting a specific pattern of looks with regard to the residue set.

If *some* is not pragmatically enriched, the residue set should not be visually consulted. The rationale behind this is as follows: upon hearing "the girl that has *some*", if the participant only accesses the literal meaning of *some* ("some and possibly all"), the linguistic input is consistent with any of the girl target regions. What is in the residue set is irrelevant. However, as the gender information has been mentioned, this could attract attention to the target regions, at the expense of the residue region. Therefore, under the slow pragmatic view, we should expect no U-shaped curve (i.e. a decrease followed by an increase) in looks to the residue set for the *some* condition. In addition, bias to residue set should be clearly less than in *all* trials.

Methods
Participants. Thirty-six participants were recruited from our university campus via an online psychological subject pool. All participants speak English as a native language. They have uncorrected or corrected to normal vision.

Materials. The experiment employed a three by two within-subject design. The two independent variables were *Determiner* (All, Some, Number) and *Target size* (Small, Big), which generated six experimental conditions: big *all*, small *all*, big *some*, small *some*, big *number* (i.e. three), small *number* (i.e. two). Experimental sentences were of the form "Click on the girl that has [determiner] of the [modifier] [shape]". [determiner] was one of *some*, *all*, *two*, *three*. [modifier] was one of *dotted*, *stripy*, *checked*; and [shape] was one of *circles*,

squares, triangles. 36 experimental displays were constructed. In each display, there were always two characters that had a total set of one kind of object and the other two had a proper subset. The residues of two partitioned sets remained in the centre. In terms of set sizes, two characters always possessed a larger set consisting of three objects, and the other two possessed a smaller set consisting of two objects.

Each display generated three experimental sentences, and three lists were created. Each list contained 36 experimental items, 6 items per condition. In addition, each list contained 18 fillers. Filler sentences were similar to experimental sentences but contained different determiners. Six fillers contained *none* (e.g. Click on the girl that has none of the objects), 6 contained *one*, and 6 contained *four* (e.g. Click on the girl that has four of the dotted circles). Twelve fillers were included to counter-balance the extra times that the target was referred by quantifiers in experimental items. Another 18 displays were constructed for fillers. Example displays of filler items are provided in Appendix B. Given that *some* and *all* were never used in Experiment 2a to refer to a set of one object or a set of four objects, the use of number terms, *one* and *four*, would not interfere with processing in a way that was discussed in Degen and Tanenhaus.

The audio descriptions and instructions were recorded in a single session by a male native British English speaker. The speaker was instructed to record all of the sentences with a neutral intonation. The audio instructions were cross-spliced in order to avoid co-articulation information in favour of any condition.[9] Across stimuli, the onset of the determiner was the same. The durations of critical time windows were adjusted using phonetics analysis software Praat (Boersma & Weenink, 2017). The average duration for the determiner time window was 773 ms (*all of the*: 741 ms, *some of the*: 793 ms, *three of the*: 784 ms, *two of the*: 773 ms), the average duration for modifier window was 596 ms (*stripy*: 597 ms, *dotted*: 594 ms, *checked*: 596 ms).

The shape (circles, triangles, squares), pattern (stripy, dotted, checked) and location of the target were counterbalanced within each condition. All pictures of an agent with a set of objects measure 336*315 pixels. Pictures of items in the middle measure 168*210. The screen resolution is 1680*1050 pixels.

Procedure. Each trial began with a display (e.g. Figure 2 (left)) in which four characters surrounded four sets of objects. Participants heard a description of the types of objects in the middle, for example, "There are stripy squares, dotted squares, stripy circles and dotted

circles". Six seconds after the onset of the description, the next display appeared (e.g. Figure 2(right)). The objects were distributed to four identical characters. After 2.5 s, participants were given an auditory instruction, for example, "Click on the girl that has some of the stripy circles". Participants' task was to click on the correct target according to the instruction. The average length of the instruction was 3.8 s. The session was set to jump to the next trial 5.5 s after the onset of the instruction. There were six practice trials in the beginning to ensure that participants understood the instruction, display and procedure. They then completed 54 trials, divided into 36 experimental trials and 18 fillers. A randomised order of presentation of the items was created for each participant.

The experiment was conducted using E-Prime software and a Tobii TX300 eye-tracker. Fixations were sampled every 17 ms. Participants were calibrated at the beginning of the experiment using a nine-point display. Before each trial, there was a fixation cross in the centre of the screen, and participants' eye gaze had to be fixed on this point for a continuous 1 s before the trial started. Eye movements were recorded from the onset of the instruction for 5.5 s for each trial. The whole experiment lasted approximately 20 min.

Data treatment and analysis methods

One per cent of the trials were excluded because participants clicked on the wrong target. Thirteen per cent of the trials were excluded due to track loss.[10] A fixation that landed within the coordinates of a character with a set of objects and the residue set was coded as a look to that area, otherwise, it was coded as background. Any fixations shorter than 80 ms were excluded, as extremely short fixations are often due to false saccade planning (Rayner & Pollatsek, 1989). Fixations were analysed in two critical time windows: the determiner window and the modifier window. The absolute onset of each time window has been offset by 200 ms for all plots and data analyses, as it takes around 200 ms to launch an eye-movement (Hallett, 1986).

We first analysed participants' eye movements to target regions (i.e. characters with objects) in the visual display. During the determiner window, the two characters that had an incomplete collection of one object type were targets for *some* trials and competitors for *all* trials, the two characters that had a complete collection were targets for *all* trials and competitors for *some* trials. During the modifier window, the target was the character of the description (e.g. the girl with some of the stripy squares), and the competitor was the character that had the objects with the same pattern (e.g. the girl with all of the stripy circles).[11]

To filter out the eye-movement based dependencies, the fixation data were aggregated over 50 ms time bins (Barr, 2008b). A target-preference score was calculated for each 50ms bin (see Arai, van Gompel, & Scheepers, 2007): $\ln(P_{(target)}/P_{(competitor)})$, where $P_{(target)}$ refers to the proportion of looks to the target, $P_{(competitor)}$ refers to the proportion of looks to the competitor, and ln refers to the natural logarithm. A score above zero indicates a greater bias towards the target and a score below zero indicates a greater bias towards the competitor. A score of zero indicates an equal bias towards the target and competitor. We fitted a linear mixed-effects model for each time window to predict target-preference scores from fixed effects of Time, Determiner, Target size and their interactions. The model contained maximal random effects structure justified by our experimental design (Barr, Levy, Scheepers, & Tily, 2013), which included random intercepts and slopes for Time, Determiner, Target size and their interaction by participants and random intercepts and slopes for Time, Determiner, Target size by items. The correlation between random intercept and random slopes was removed. The 3-level factor Determiner and the 2-level factor Target size were deviation coded, and the continuous factor Time was centred. Model comparisons were conducted to test the significance of fixed effects with more than two levels, using likelihood ratio tests. Significant main effects were followed up using re-referencing approach,[12] and significant interactions were followed up by conducting analyses on the subset of the data, only including the relevant pair of conditions.

We then investigated participants' eye movements to the residue set. Growth curve analyses were conducted to capture the rise and fall in fixations to the residue set over the determiner window and the modifier window (Mirman, 2017; Mirman, Dixon, & Magnuson, 2008). For the growth curve analysis, the empirical logit[13] was calculated for each 50ms bin, which is a quasi-logit transformation of fixation probability (i.e. looks to the residue set over looks to other areas) (Barr, 2008b). We fitted a model to predict empirical logits from fixed effects of Time, Determiner, Target size and their interactions. Time was a continuous variable represented by Time 1 and Time 2. Time 1 is the linear representation of Time, and Time 2 is the quadratic representation of Time (2nd-order orthogonal polynomial). The 3-level factor Determiner was treatment coded with number as the reference level, and the 2-level factor Target size was deviation coded (Small, −0.5; Big, 0.5). The model included random intercepts for participants and items. All fixed effects were included as random slopes for participants and items, interactions were not included as random slopes because the model failed to converge. The correlation between random intercept and random slopes was removed.

All statistical analyses were carried out using the lme4 package (Bates, Maechler, Bolker, & Walker, 2015) in R (R Core Team, 2017). For both linear mixed-effect models and growth-curve analyses, we report the regression coefficient, the standard error, and t-value. Significances of effects were assessed by checking whether the absolute value of the t-statistic exceeds 2 (Baayen, 2008). Below we first report analyses of eye movements toward target regions. We then report analyses of eye movements toward the "residue set".

Analyses of eye movements towards target regions

Given that areas of interest changed as the sentence unfolded, we visualised how target preference developed in two critical time windows separately, as shown in Figures 4 and 5. In these graphs, we plot the average target preference score for each condition for every 17 ms sample. With regard to the experimental sentence, "Click on the girl that has some of the stripy squares", Figure 4 covers the time region from the sentence onset to the offset of the determiner window (i.e. "the" offset), and Figure 5 covers from the onset of the modifier window "stripy" to the end of the sentence. Curves in these log-ratio plots were resynchronised at each time window onset to ensure that the target bias formation was visually represented accurately (Altmann & Kamide, 2009).

We were interested in two questions in the following analyses: (i) whether the hypothesised factors influence target preference for the all and the some condition, and (ii) how target preference differs across all, some and number conditions. To answer these questions, we constructed separate linear mixed-effects models for the determiner window and the modifier window predicting target preference scores from fixed effects of Determiner (all, some, or numbers), Target size (small or big), Time and their interactions, including random effects structure as described above. In the following we will first report the results of the determiner window, then the results of the modifier window.

Determiner window. Figure 4 (top) depicts how target preference developed from the instruction onset to the offset of the determiner window by determiner type. First, we found a main effect of determiner type ($\chi^2(2) = 10.27$, $p = .006$). Post hoc analyses revealed that on average the target preference was significantly greater in the number condition compared to the all condition ($\beta = -0.56$, SE = 0.18, $t = -3.07$) and the some condition ($\beta = -0.58$, SE = 0.18, $t = -3.27$), and there was no
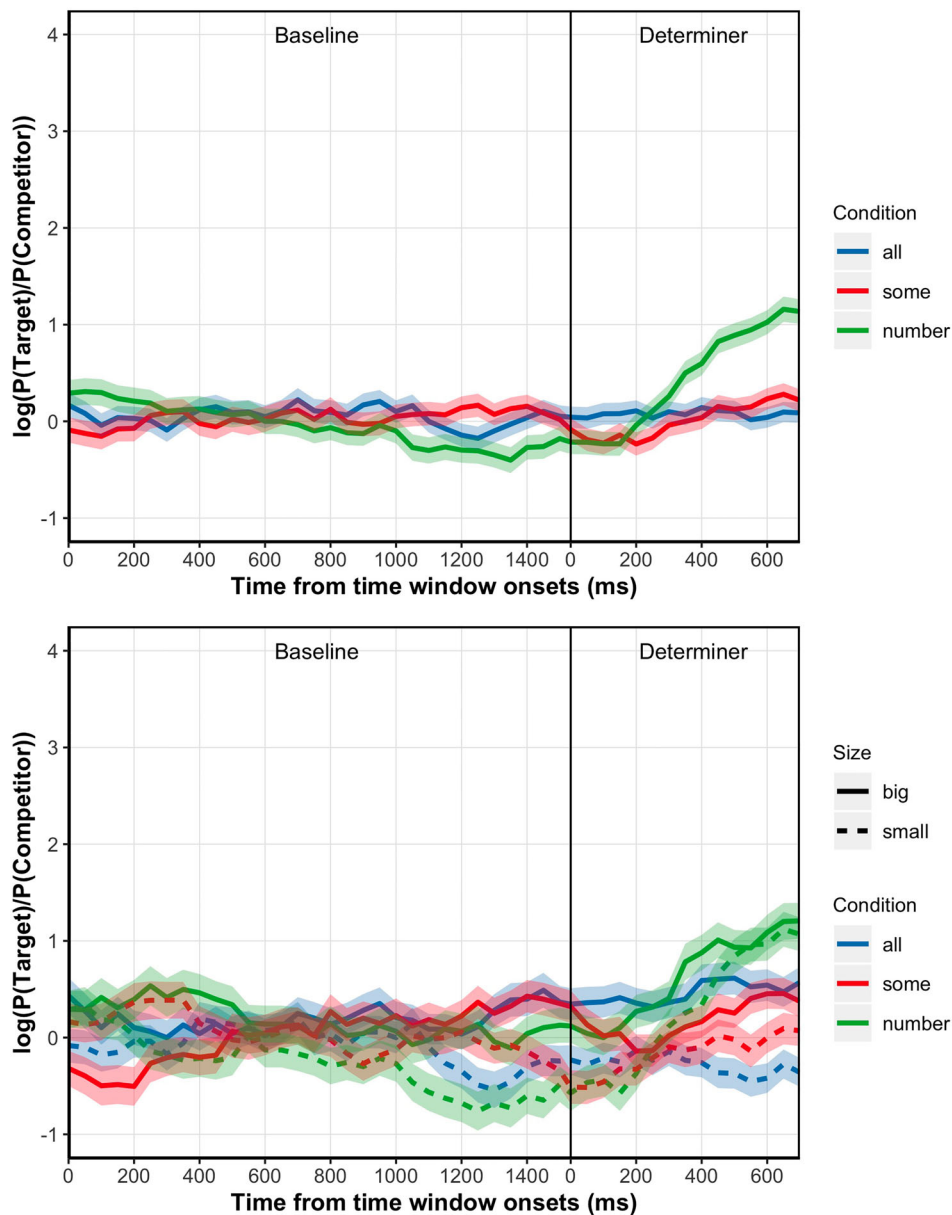
Figure 4. Target preference scores from the instruction onset to the determiner window offset in Experiment 2a. The top graph shows target preference scores by determiner type, and the bottom graph shows target preference scores by determiner type and target set size. Time 0 ms indicates the onset of each time window. Standard errors are represented by transparent ribbons.

significant difference between the *all* and *some* conditions ($t = -0.07$, ns). There was also a main effect of time ($\beta = 1.21$, SE = 0.33, $t = 3.63$) and a significant interaction between determiner type and time ($\chi^2(2) = 14.02$, $p < .001$). Post hoc analyses revealed that an increased target preference over time was only present in the *number* condition ($\beta = 2.76$, SE = 0.53, $t = 5.24$), but not in the *all* or *some* condition ($ts < 2$). Between *all* and *some*, target preferences did not develop at different rates ($t = 1.56$, ns).

Figure 4(bottom) depicts how target preference developed over time by determiner type and target size. We found a main effect of target size ($\beta = 0.56$, SE = 0.21, $t =$

2.63), with significant greater target preference in the big-set target condition ($M = 0.56$, SD = 3.25) compared to the small-set target condition ($M = 0.01$, SD = 3.33). Analyses did not reveal any significant interaction between target size and time ($t = -1.18$, ns) or determiner type ($\chi^2(2) = 2.79$, ns). However, there was a marginal significant three-way interaction between target size, determiner type and time ($\chi^2(2) = 5.82$, $p = .05$).

To understand this three-way interaction, we first look at the simple effect of target size within each determiner type. We found that the effect of target size was only present in the *all* condition ($\beta = 0.95$, SE = 0.35, $t = 2.73$), where the target preference was greater when *all* was
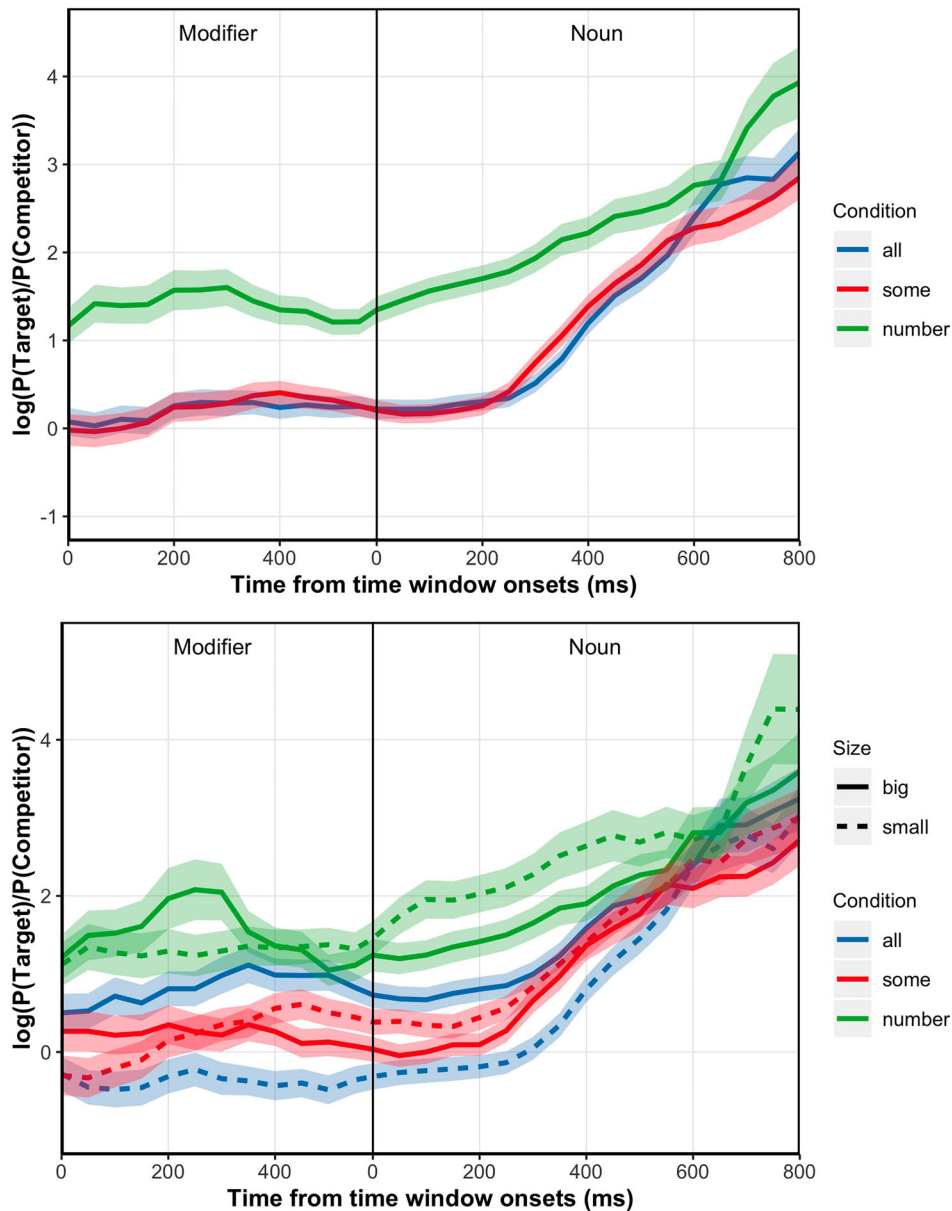
Figure 5. Target preference scores from the modifier onset to the instruction offset in Experiment 2a. The top graph shows target preference scores by determiner type, and the bottom graph shows target preference scores by determiner type and target set size. Standard errors are represented by transparent ribbons.

used with a larger set ($M = 0.61$, SD $= 3.28$) compared to when it was used with a smaller set ($M = -0.36$, SD $= 3.37$). There was no effect of target size in the *some* or *number* conditions ($|t|$s < 2). We then turn to look at the simple effect of determiner type and the determiner type by time interaction on each level of target size. When the target was a larger set, there was no difference among the *all*, *some* and *number* conditions on either the average target preference score or the rate of increase of target preference ($|t|$s < 2). The story was different when the target was a smaller set. The target preference was significantly greater and increased faster in the *number* condition compared to the *all* and *some*

conditions (*all*: $\beta = -4.29$, SE $= 0.71$, $t = -6.03$; *some*: $\beta = -1.96$, SE $= 0.56$, $t = -3.5$). Crucially, while the overall target preference did not differ between *some* and *all*, preference increased faster in the *some* condition than in the *all* condition ($\beta = 2.47$, SE $= 0.69$, $t = 3.59$).

Modifier window. Figure 5 depicts how target preference developed by determiner type and by determiner type and target size from the onset of the modifier window to the instruction offset. Again we found a main effect of determiner type ($\chi^2(2) = 44.6$, $p < .05$). Similar to the effect observed in the determiner window, on average the target preference was

significantly greater in the *number* condition compared to the *all* condition ($\beta = -1.38$, SE $= 0.25$, $t = -5.49$) and the *some* condition ($\beta = -1.31$, SE $= 0.24$, $t = -5.4$), and there was no significant difference between the *all* and *some* conditions ($t = 1.07$, ns). In the modifier window, neither the main effect of time ($t = 1.28$, ns) nor the interaction between time and determiner type ($\chi^2(2) = 0.18$, ns) was significant.

The main effect of target size ($\beta = 0.53$, SE $= 0.23$, $t = 2.34$) continued, with significant greater target preference in the big-set target condition ($M = 1.05$, SD $= 3.37$) compared to the small-set target condition ($M = 0.49$, SD $= 3.52$). Whilst there was no significant interaction between target size and time ($t = -0.99$, ns), analyses revealed a significant interaction between target size and determiner type ($\chi^2(2) = 11.39$, $p = .003$).

To understand this two-way interaction, we first consider the effect of target size on each level of determiner type, and then consider the effect of determiner type on each level of target size. We found that, similar to the determiner window, the effect of target size was only present in the *all* condition ($\beta = 1.58$, SE $= 0.37$, $t = 4.23$), where the target preference was greater when *all* was used with a larger set ($M = 1.02$, SD $= 3.35$) compared to when it was used with a smaller set ($M = -0.54$, SD $= 3.42$). No effect of target size was found in the *some* or *number* conditions ($|t|$s < 2). As for the effect of determiner type, we found that when the target was a larger set, the target preference was significantly greater in the *all* condition compared to the *some* condition ($\beta = -0.65$, SE $= 0.25$, $t = -2.59$). Whereas when the target was a smaller set, the effect flipped such that the target preference was greater in the *some* condition compared to the *all* condition ($\beta = 0.88$, SE $= 0.32$, $t = 2.79$). In addition, the target preference was always significantly greater in the *number* condition compared to both the *all* and *some* conditions.[14] There was no significant three-way interaction between time, determiner and target size ($\chi^2(2) = 0.02$, ns).

Discussion of target region analyses. Table 4 provides a summary of the results, focusing on the influence of set size on target preference for the *all* and the *some*

condition. In both time windows the target bias in the *all* condition was greater when the target was a larger set compared to when it was a smaller set. In addition, as can be seen more clearly in the modifier window, the target bias was greater in the *all* condition compared to the *some* condition when the target was a larger set, but it was smaller in the *all* condition when the target was a smaller set. These results confirmed the predictions that a combination of the low-level association between *all* and the larger set plus Maximise Presupposition influence the target preference for the *all* condition in a way that the target bias was boosted when the target was a larger set and was disadvantaged when the target was a smaller set.

In Experiment 2a, target biases in the *some* condition did not differ between different set sizes. Thus, although Experiment 1a suggested that the use of *some* might trigger a low-level preference for the larger set, this offline preference was not detected.

Comparing the overall target preference for *number* vs. *all* and *some*, we found that the *number* condition showed a steeper increase in target bias over the determiner window and a greater overall target bias in the modifier window, compared to the *all* and *some* conditions. These results are in line with our predictions drawn from differences in verification procedures among conditions. In addition, we compared target biases in the *all* and *number* conditions when the target was a set of three objects. This condition pair is comparable with the condition pair of *all* and *three* in Huang and Snedeker (2009). In the determiner window analyses, we replicated the no-difference result reported in Huang and Snedeker (2009), whereas in the modifier window analyses we found a greater target bias in the *number* condition. Focusing on the comparison in the determiner window, Huang & Snedeker explained the no-difference result by suggesting that target identification in the *all* and *number* conditions only relies on the integration of the literal meaning. However, this explanation ignores the fact that verifying *all* should take longer than *numbers*. Thus, instead we explain this no-difference finding in terms of other factors offsetting the cost of more complex verification on *all*.

Table 4. Summary of pairwise comparisons in Experiment 2a, 2b and 3.

	Experiment 2a		Experiment 2b		Experiment 3	
	Determiner window	Modifier window	Determiner window	Modifier window	Determiner window	Name window
all	**big-set > small-set**	**big-set > small-set**	big-set = small-set	**big-set > small-set**	**big-set > small-set**	**big-set > small-set**
some	big-set = small-set	big-set = small-set	big-set = small-set	big-set = small-set	big-set = small-set	big-set = small-set
big-set	all = some	***all > some***	all = some	***all > some***	all = some	***all > some***
small-set	**all < some***	**all < some**	all = some	**all < some**	all = some	all = some

Notes: > indicates that the target preference is greater in one condition than the other, and < indicates the target preference is smaller in one condition than the other. = indicates no significant effect. * indicates the presence of the effect on the linear term of *Time*. *italics* indicates that the effect is predicted by *Slow Pragmatic*, **bold** indicates that the effect is predicted by our account, and bold ***italics*** indicates that the effect is predicted by both accounts.

Comparing the overall target preference for *all* vs. *some*, we found no difference in target preference between the *some* and *all* conditions in both determiner and modifier windows, even with the use of number words. This finding contrasts with previous reports of delayed *some* (Huang & Snedeker, 2009; 2011; 2018). An important difference between these previous studies and Experiment 2a is that the target set size for the *all* and *some* referents were fully counterbalanced in the latter. Thus, the conflicting finding might be due to this difference, and Experiment 2a provides a compelling piece of evidence for the fast pragmatic view. An alternative interpretation of the results could concede that a low-level association and Maximise Presupposition might have been a factor in previous studies but that, on top of these factors, there is nevertheless an underlying delay on *some* trials as against *all*.[15] For example, results in the modifier window show greater bias for big-set *all* compared to big-set *some*. The question is whether this difference on big-set trials is solely driven by the factors we have identified or not. We think that the results in small-set comparisons, where the difference between *all* and *some* is reversed in the manner we predict would cast doubt on this alternative account. For if there is an underlying delay for *some* relative to *all*, then the reversal of difference in small-set trials should be difficult to detect. In fact, we detected this difference in both time windows.

Our results are not identical, but comparable to those in the small set-size condition of Degen and Tanenhaus

(2016), which included only set sizes two and three, while counterbalancing size with determiner.

Analyses of visual search to the residue set

Figure 6 shows the rise and fall in fixations to the residue set in the *some*, *all* and *number* conditions from the sentence onset to the end. In the following analyses, we were interested in two questions: (i) is the overall bias towards the residue set in the *all* and *some* conditions greater than in the *number* condition, and (ii) is there a U-shape parabolic change in fixation probability on the residue set in the *all* and *some* but not in the number condition?

We constructed separate growth curve analyses for the determiner window and the modifier window predicting the empirical logit of fixation probabilities from fixed effects of *Determiner*, *Target size*, *Time* and their interactions. *Time* was represented by *Time 1* and *Time 2* to capture how the probability of fixating the residue set changed over time. *Time 1* modelled a straight line that has an initial change of direction from flatness; *Time 2* modelled a U-shape curve that has an initial change from flatness and the reversal at the bottom of U (Mirman, 2017). Following this, in our visual-world study, a positive coefficient of *Time 1* indicates an increase in fixations to the residue set over time, and a negative coefficient indicates a decrease in residue set fixations. As for the coefficient of *Time 2*, a positive coefficient indicates an initial decrease in fixations followed by an increase in fixations to the residue set, whereas a negative coefficient of *Time 2* would just be inverted.
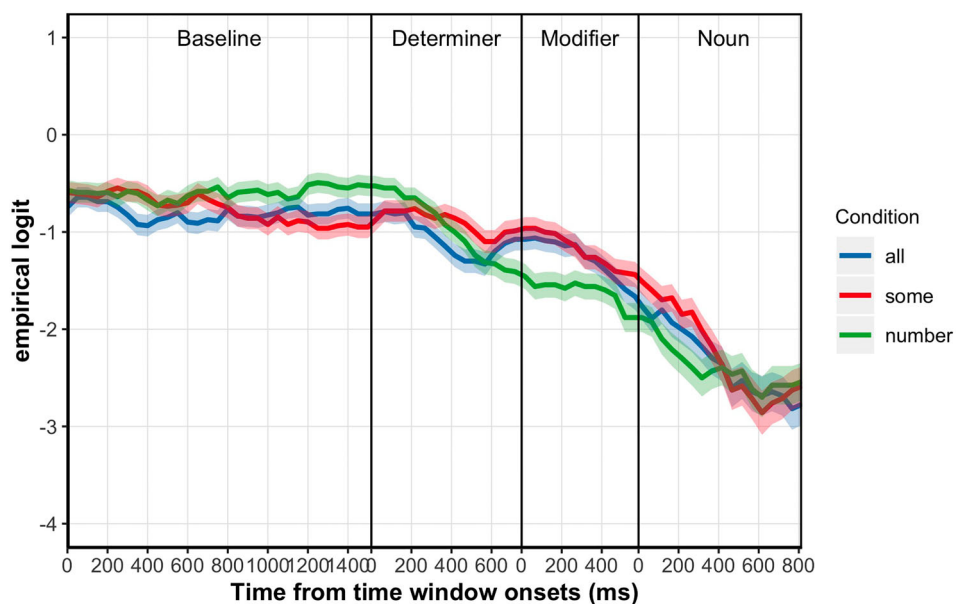

Figure 6. Fixation probabilities (empirical logit transformation) on the residue set by determiner type from the instruction onset to the instruction offset in Experiment 2a. Standard errors are represented by transparent ribbons.

Determiner window. We found a significant main effect of *Time 1* ($\beta = -0.42$, SE $= 0.14$, $t = -2.93$) and a significant interaction between determiner type and *Time 1* ($\chi^2(2) = 54.36$, $p < .001$). Post hoc analyses revealed that fixations on the residue set was decreased linearly only in the *number* condition ($\beta = -0.82$, SE $= 0.20$, $t = -4.18$), but not in the *some* or *all* conditions ($|t|$s < 2). We also found a main effect of *Time 2* ($\beta = 0.28$, SE $= 0.07$, $t = 3.78$) and a significant interaction between determiner type and *Time 2* ($\chi^2(2) = 12.37$, $p = .002$). Post hoc analyses revealed a significant upward curving quadratic component for the *all* condition only ($\beta = 0.47$, SE $= 0.13$, $t = 3.72$). Coefficients of *Time 2* were positive in the *some* and *number* conditions but did not reach significance (*some*: $\beta = 0.11$, SE $= 0.10$, $t = 1.13$; *number*: $\beta = 0.27$, SE $= 0.14$, $t = 1.95$). In addition, there was a significant interaction between target size and *Time 2* ($\beta = 0.26$, SE $= 0.08$, $t = 3.20$), with greater curvature in looks to the residue set when the target was the larger set compared to when it was the smaller set. Other main effects and interactions were not significant.

Modifier window. We found a significant main effect of determiner type ($\chi^2(2) = 6.58$, $p = .04$). Post hoc analyses revealed that on average fixations on the residue set were significantly greater in the *all* condition and the *some* condition compared to the *number* condition (*all*: $\beta = 0.31$, SE $= 0.10$, $t = 3.01$; *some*: $\beta = -0.31$, SE $= 0.11$, $t = 2.79$), and there was no significant difference between the *all* and *some* conditions ($t = 0.13$, n.s.). The main effect of *Time 1* continued ($\beta = -0.36$, SE $= 0.09$, $t = -4.11$), but there was no significant interaction between determiner type and *Time 1* ($\chi^2(2) = 0.61$, ns).

We found a significant main effect of *Time 2* ($\beta = 0.25$, SE $= 0.09$, $t = 2.96$), and a significant interaction between determiner type and *Time 2* ($\chi^2(2) = 7$, $p = .03$). Post hoc analyses revealed a significant effect of *Time 2* only for the *all* and *some* conditions (all: $\beta = 0.27$, SE $= 0.12$, $t = 2.25$; some: $\beta = 0.37$, SE $= 0.11$, $t = 3.31$), not for the *number* condition ($t = 1.09$, ns). Between the *all* and *some* conditions, there was no difference in the quadratic curvature ($t = 1.03$, n.s.). In addition, similar to the determiner window, there was a significant interaction between *Time 2* and target size ($\beta = 0.16$, SE $= 0.07$, $t = 2.20$), with greater curvature in looks to the residue set when the target set was the larger set compared to when it was the smaller set. Other main effects and interactions were not significant.

Discussion of residue set analyses. In the determiner window, we found the overall bias towards the residue set differed little among determiners. However, there

was an effect of *Determiner* on both the linear and quadratic terms of *Time*. On the linear term of *Time*, only the *number* condition revealed a decrease in fixations towards the residue set; on the quadratic term of *Time*, only the *all* condition revealed a U-shape parabolic change in fixation probability on the residue set. In the modifier window, the overall bias towards the residue set had become greater in the *all* and *some* conditions compared the *number* condition. There was an effect of *Determiner* on the quadratic, not linear, term of *Time*. We found a U-shape pattern in fixation probability in the *all* and the *some* condition, but not in the *number* condition. The U-shape pattern indicated that the initial decrease in residue set fixations was followed by a latter increase, which participants shifted their fixations from other areas in the display to the residue set.

Both the fast pragmatic and the slow pragmatic view predict the difference in visual biases between the *all* and the *number* condition. However, only the fast pragmatic view predicts the difference in visual biases between the *some* and the *number* condition. According to the fast pragmatic view, the rapid integration of pragmatic *some* would draw participants' visual attention to the residue set while determining the correct target. This would not only lead to a greater visual bias in the *some* condition compared to the *number* condition, but would also cause two changes of direction in the fixation data in the *some* condition. The data from the residue set analyses, therefore, is in line with these predictions, and provides a novel piece of evidence confirming the fast pragmatic view.

In both time windows, we found the overall bias towards the residue set was not affected by target size, suggesting that visual biases to the residue set were less affected by target set size compared to biases to the target region. We observed an effect of *Target size* on the quadratic term of *Time*, indicating that the U-shaped curve was sharper when the target was a larger set compared to when it was a smaller set. It is not clear how to interpret this pattern since there was no interaction with determiners. As the time increased, the increased difference between the U-shape curves for big set and small set indicates that participants' rate of attention shifts (goes first to the target area and then to the residue area) was greater when the target was a larger set compared to when it was a smaller set.

In both time windows, we found no significant U-shape parabolic change in the *number* condition. This finding again reflects the difference in verification processes between numbers and non-numbers. Note that although the effect was not significant, we found in the *number* condition, some "return" to the residue set during the determiner window. This is not predicted on any

account of the online interpretation of these items. We suspect that shifts to the residue set in the *number* condition were due to the use of subitisable sets. When the number of objects in a set is within subitisable range (1-3), the quantity of the set is rapidly available and salient to participants through pre-attentive visual recognition processes (Dehaene, 1997). Thus, in the *number* condition, after identifying two referents that had the correct amount of objects, participants might have time to fixate any viewing region before the modifier onset. The shifts toward the residue set might reflect such a "noise" event. If our suspicion is correct, the "return" effect in number trials is less likely to occur in the next experiment, where the targets may be less straightforward to recognise and distinguish, since one of the numbers, *four*, is on the outer edge of the subitisable region while the pair (3/4) involve a smaller difference (larger Weber fraction) and are less discriminable than the pair (2/3).

Experiment 2b

Similar to Experiment 2a, each trial started with either Figure 7 (left) or Figure 8 (left) followed by Figure 7 (right) or Figure 8 (right). Participants were asked to click on the referent of the audio instruction, e.g. "Click on the girl that has *some* of the stripy squares". We again manipulated the target set size for the *all* and the *some* referent. In contrast to Experiment 2a, larger sets in the display contained four objects, and smaller sets contained three objects. Based on the results of Experiment 1b, we expect that after hearing *all*, participants' anticipatory looks should be biased towards the larger sets in the display. In other words, differences in target biases between the condition pair big-set *all* and small-set *all* found in Experiment 2a should be replicated.

In addition, we no longer expect to see a greater target bias in big-set *some* compared to small-set *some*. Like Experiment 2a, number items were included. As was discussed, using the number *four*, to some extent, controls for the subitisability. We expect to no longer see "return" looks to the residue set in the *number* condition during the determiner window.

Methods
Participants. Thirty-six participants were recruited from our university campus via an online psychological subject pool. All participants speak English as a native language. They have uncorrected or corrected to normal vision.

Procedure and materials. The materials were the same as Experiment 2a except that in Experiment 2b we changed the target set sizes. In experimental displays, the larger sets contained four objects and the smaller sets contained three objects. Thus, determiners used in the *number* condition were changed to *three* and *four*. As in Experiment 2a, three lists were created. Each list contained 36 experimental items, 6 items per condition. Again, there were 18 fillers, of which 12 were number fillers. The determiner used in the filler item was one of *none*, *two* and *one*. The audio instructions were cross-spliced and adjusted. The average length of the instruction was 4.1 s. The average duration for the determiner window was 718 ms (*all of the*: 703ms, *some of the*: 725ms, *four of the*: 720 ms, *three of the*: 729 ms), and the average duration for the modifier window was 632ms (*stripy*: 638 ms, *dotted*: 638 ms, *checked*: 622 ms). The procedure was identical to Experiment 2a except for one difference. That is, given the complexity of the display, participants were given more time to respond
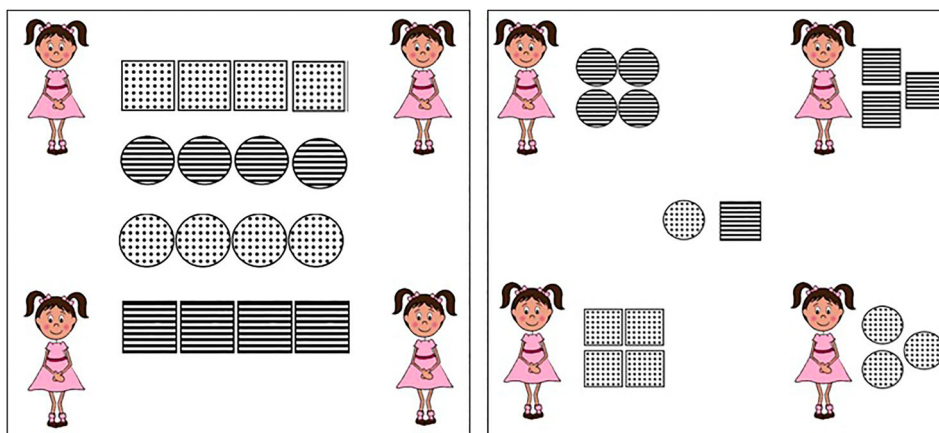

Figure 7. Example displays for Experiment 2b: big *all* / small *some*. The left image is the initial display, and the right image is the critical display. Figure 7 (right) can be paired with "Click on the girl that has all/four of the stripy circles" or "Click on the girl that has some/three of the stripy squares".
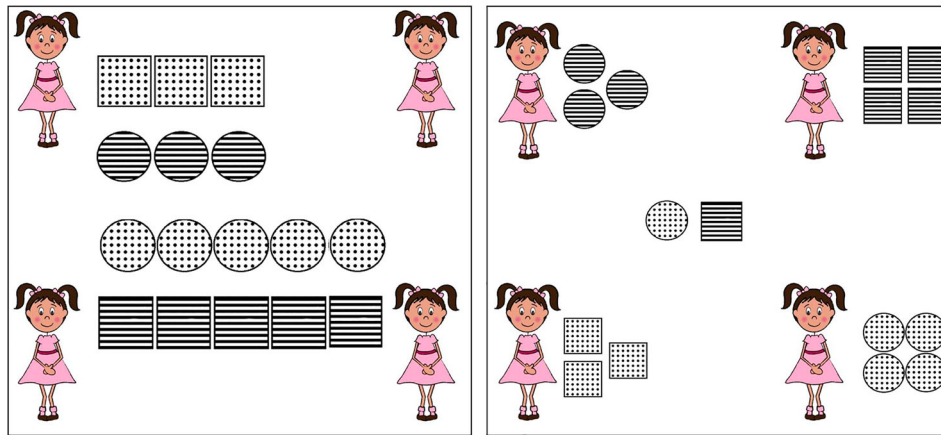
Figure 8. Example displays for Experiment 2b: small *all* / big *some*. The left image is the initial display, and the right image is the critical display. Figure 8 (right) can be paired with "Click on the girl that has all/three of the stripy circles" or "Click on the girl that has some/four of the stripy squares".

in each trial. It was set to jump to the next trial 6 s after the instruction onset.

Data treatment

Two per cent of the trials were excluded because participants clicked on the wrong target. Twenty-eight per cent of the trials were excluded due to track loss. Again, we first analysed participants' eye movements to target regions, we then investigated participants' eye movements to the residue set.

Analyses of eye movements towards target regions

As in Experiment 2a, we defined two critical time windows: the determiner window and the modifier window. The target(s) and competitor(s) in each window were also defined in the same way as in Experiment 2a. Figures 9 and 10 visualised how target preference developed in these two windows separately. Similarly, we constructed separate linear mixed-effects models for each time window predicting target preference scores from fixed effects of *Determiner* (*all*, *some*, or *numbers*), *Target size* (small or big), *Time* and their interactions, including random effects structure as described above. In the following we will first report the results of the determiner window, then the results of the modifier window.

Determiner window.
We found a significant main effect of time ($\beta = 1.40$, SE $= 0.33$, $t = 4.22$) and a significant interaction between time and determiner type ($\chi^2(2) = 14.62$, $p < .001$). Post hoc analyses revealed that the effect of time was only present in the *number* condition ($\beta = 3.16$, SE $= 0.51$, $t = 6.17$), but not in the *all* or *some* condition ($ts < 2$). Between *all* and *some*, target preferences did not develop at different rates ($t = 0.001$, ns).

In contrast to Experiment 2a, other main effects and interactions were not significant.

Modifier window.
We found a main effect of determiner type ($\chi^2(2) = 17.07$, $p < .001$). On average the target preference was significantly greater in the *number* condition compared to the *all* and *some* conditions (all: $\beta = -1.20$, SE $= 0.26$, $t = -4.61$; some: $\beta = -1.38$, SE $= 0.26$, $t = -5.27$), and the *all* and *some* conditions did not differ from each other ($t = -0.83$, ns). We also found a significant interaction between target size and determiner type ($\chi^2(2) = 13.52$, $p < .001$). To understand this two-way interaction, we first consider the effect of target size within each determiner type. We found that the effect of target size was present in the *all* condition ($\beta = 1.08$, SE $= 0.46$, $t = 2.35$), where the target preference was greater when *all* was used with a larger set ($M = 1.20$, SD $= 3.41$) compared to when it was used with a smaller set ($M = -0.01$, SD $= 3.42$). No effect of target size was found in the *some* condition. Somewhat unexpected, there was an effect of target size in the *number* condition, with greater target bias for *four* than for *three* ($\beta = 0.74$, SE $= 0.32$, $t = 2.32$). We then turn to look at the effect of determiner type on each level of target size. We found that when the target was a larger set, the target preference was significantly greater in the *all* condition compared to the *some* condition ($\beta = -1.02$, SE $= 0.28$, $t = -3.59$). Whereas when the target was a small set, the effect flipped such that the target preference was greater in the *some* condition compared to the *all* condition ($\beta = 0.71$, SE $= 0.34$, $t = 2.11$). The target preference was always significantly greater in the *number* condition compared to both the *all* and *some* conditions.[16] There was also a significant three-way interaction
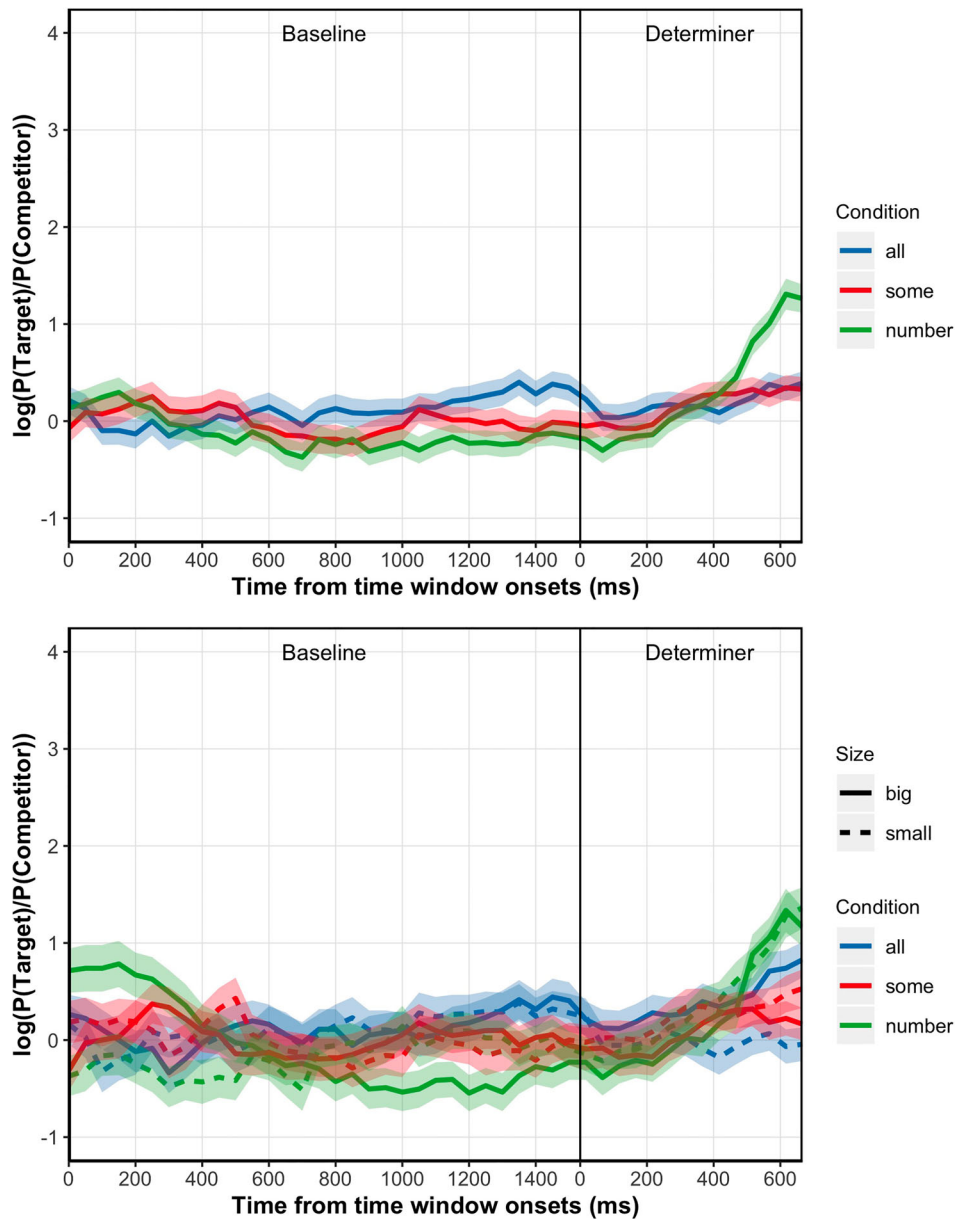
Figure 9. Target preference scores from the instruction onset to the determiner window offset in Experiment 2b. The top graph shows target preference scores by determiner type, and the bottom graph shows target preference scores by determiner type and target set size. Standard errors are represented by transparent ribbons.

between determiner type, target size and time ($\chi^2(2) = 7.86$, $p = .02$). This interaction was driven by a steeper increase in the target bias of the *number* condition compared with that of the *all* and *some* conditions when the target was a smaller set. Other main effects and interactions were not significant in the modifier window.

Discussion of target region analyses. As summarised in Table 4, there was little difference between conditions in the determiner window, but in the modifier window Experiment 2b replicated the finding of Experiment 2a. These results confirmed that first, the low-level

association between *all* and the larger set affects eye movements in the predicted way; second, the association, if any, between *some* and set size does not affect online measures.

Regarding the comparison of the overall target bias among determiners, Experiment 2b replicated the findings of Experiment 2a. The target bias of the *number* condition showed a steeper increase over the determiner window and a greater overall in the modifier window, compared to the *all* and *some* conditions. Also there was no difference in target preference between the *some* and *all* conditions in both time
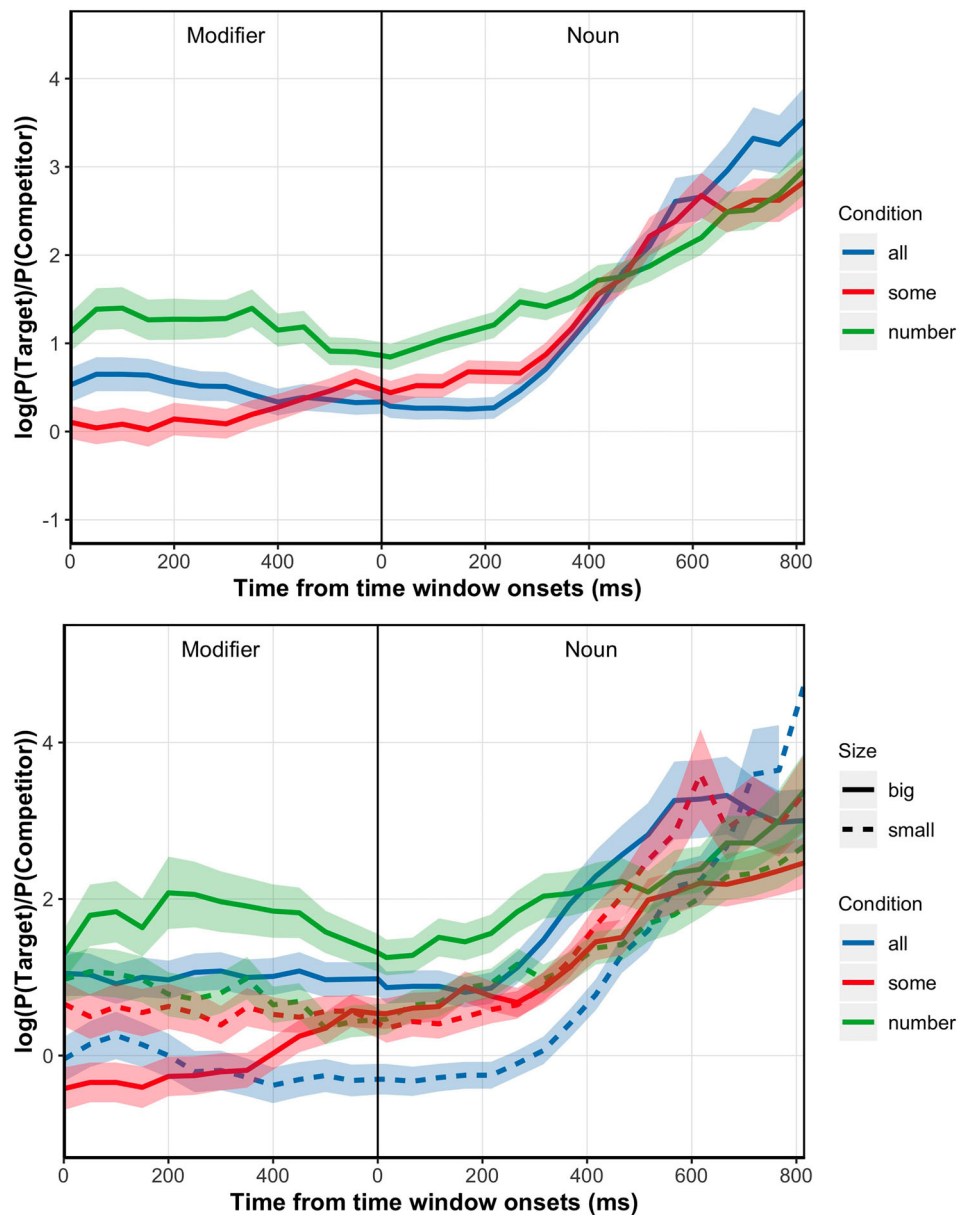
Figure 10. Target preference scores from the modifier onset to the instruction offset in Experiment 2b. The top graph shows target preference scores by determiner type, and the bottom graph shows target preference scores by determiner type and target set size. Standard errors are represented by transparent ribbons.

windows. Thus, Experiment 2b provides further data that support the fast pragmatic view.

Analyses of visual search to the residue set

Figure 11 depicts participants' visual search to the residue set over time by determiner type. As we did for Experiment 2a, we constructed separate growth curve analyses for the determiner window and the modifier window predicting the empirical logit of fixation probabilities from fixed effects of *Determiner*, *Target size*, *Time* and their interactions. *Time* was represented by *Time 1* and *Time 2* to capture both the linear and quadratic change in fixations over time.

Determiner window. We found a significant main effect of *Time 1* ($\beta = -0.55$, SE = 0.13, $t = -4.18$), and a significant interaction between *Time 1* and determiner type ($\chi^2(2) = 37.13$, $p < .001$). Post hoc analyses revealed that fixations on the residue set decreased linearly in the *number* condition ($\beta = -0.96$, SE = 0.19, $t = -4.99$) and the *all* condition ($\beta = -0.48$, SE = 0.21, $t = -2.28$), but not in the *some* condition ($t = -0.66$, ns). The *number* condition showed a steeper decrease compared to the *all* condition ($\beta = 0.33$, SE = 0.11, $t = 3.01$). We found no main effect of *Time 2* ($t = 0.75$, n.s.) but a significant interaction between determiner type and *Time 2* ($\chi^2(2) = 21.17$, $p < .001$). Post hoc analyses revealed greater
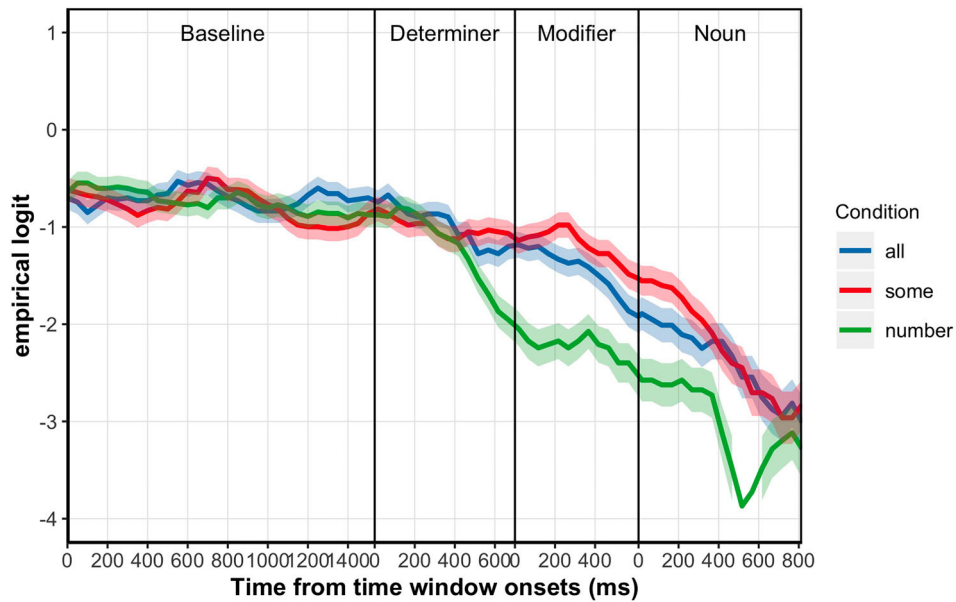
Figure 11. Fixation probabilities (empirical logit transformation) on the residue set by determiner type from the instruction onset to the instruction offset in Experiment 2b. Standard errors are represented by transparent ribbons.

curvature in looks to the residue set in the *all* and *some* conditions compared to the *number* condition (*all*: $\beta = 0.46$, SE = 0.11, $t = 4.21$; *some*: $\beta = 0.39$, SE = 0.11, $t = 3.60$). Other main effects and interactions were not significant.

Modifier window. We found a main effect of determiner type ($\chi^2(2) = 9$, $p = .01$). Post hoc analyses revealed that the average fixations on the residue set was greater in the *all* and *some* conditions compared to the *number* condition (*all*: $\beta = 0.32$, SE = 0.10, $t = 3.14$; *some*: $\beta = 0.38$, SE = 0.11, $t = 3.36$), and there was no significant difference between the *all* and *some* conditions. The main effect of *Time 1* continued ($\beta = -0.27$, SE = 0.09, $t = -2.88$), and there was a significant interaction between determiner type and *Time 1* ($\chi^2(2) = 19.62$, $p < .001$). Post hoc analyses revealed that fixations on the residue set was decreased linearly in the *all* and *some* conditions (*all*: $\beta = -0.42$, SE = 0.16, $t = -2.60$; *some*: $\beta = -0.39$, SE = 0.18, $t = -2.17$), but not in the *number* condition ($t = 0.17$, ns). Furthermore, the *all* and *some* conditions did not differ on the linear slope.

We found a significant main effect of *Time 2* ($\beta = 0.17$, SE = 0.08, $t = 2.15$), but there was no significant interaction between determiner type and *Time 2*. We found a significant interaction between target size and *Time 2* ($\beta = 0.18$, SE = 0.08, $t = 2.21$), and a significant three-way interaction between determiner type, target size and *Time 2* ($\chi^2(2) = 10.07$, $p = .007$). Post hoc analysis revealed that only for the *all* condition, there was a greater curvature in fixations towards the residue set when the target

was a larger set compared to when it was a smaller set ($\beta = 0.55$, SE = 0.19, $t = 2.95$). Other main effects and interactions were not significant. We attribute the greater curvature for big-set *all* in the later time window to a delay in search of the residue initiated on the basis of the compositional meaning of *all* after greater initial target search on the basis of the low-level association.

Discussion of residue set analyses. In the determiner window, like Experiment 2a, we found no difference in overall bias among conditions, and there was an effect of *Determiner* on both the linear and quadratic terms of *Time*. However, unlike Experiment 2a, on the linear term of *Time*, we found both the *number* and the *all* condition revealed a decrease in fixations towards the residue set, though fixations decreased faster in the *number* condition. By contrast, there was no significant decrease for *some*. Also unlike Experiment 2a, on the quadratic term of *Time*, we found greater curvature in both the *all* and *some* conditions compared to the *number* condition.

In the modifier window, as predicted we found that the overall bias towards the residue set had become greater in the *all* and *some* conditions compared the *number* condition. However, there was only a main effect of *Time 2*, suggesting fixations to the residue set decreased first and increased a little over this time window regardless of the determiner type.

Although no significant positive coefficient of *Time* 2 was found in the *all* and the *some* condition, we did find a positive quadratic trend in *some* and *all* in both

determiner and modifier window. We thus feel confident in attributing these effects to participants searching the residue set to determine which girl has *some and not all* and which girl has *all* of the relevant targets. Taken together with the finding that the *some* condition showed greater curvature compared to the *number* condition in the determiner window and greater bias in the modifier window, the residue set analyses in Experiment 2b tend to disconfirm the slow-pragmatic account, suggesting that the "some and not all" interpretation of *some* is accessed rapidly.

Summary of Experiments 2a and 2b

The findings of Experiments 2a and 2b provide supporting evidence that the low-level association between *all* and the larger set influences participants' visual attention to the target. In addition, it is possible that in Experiment 2a, the tendency to favour the larger set in *all* trials was increased by a dis-preference to use *all* with a set size specifically of two (Maximise Presupposition). Experiments 2a,b showed that when set size was controlled, the time course of target identification for the *all* and the *some* condition did not differ. In contrast to previous visual world studies, the design of Experiments 2a,b included a DV of looks to residue set.[17] The greater visual bias towards the residue set in the *all* condition compared to the *number* condition, together with a U-shaped curve in residue set fixations in *all*, reflect different verification procedures required by the *all* and the *number* condition. Moreover, the greater visual bias towards the residue set in the *some* condition compared to the *number* condition, together with a U-shaped curve similar to the curve in *all*, provide a novel piece of evidence for the fast pragmatic view.

However, we note that in Experiment 2, for both quantifiers, target identification was relatively slow, especially compared with previous studies. To determine whether the target referent was identified before the noun onset, we performed one-sample t-test to compare target proportions to chance (50%) over the modifier window. We found that, in Experiment 2a, the target proportion for *number* conditions was significantly above chance $(t_1(35) = 4.70, p < .001; t_2(35) = 6.66, p < .001)$, whereas bias in neither *all* nor *some* conditions was significantly above chance. In Experiment 2b, the target proportion for *number* and *all* conditions was significantly above chance (number: $t_1(34) = 7.38, p < .001; t_2(35) = 7.35, p < .001$; all: $t_1(34) = 3.10, p = .004; t_2(35) = 2.39, p = .02$), whereas *some* condition was still not significantly above chance.

We attribute the slow target identification to a number of mitigating factors. Compared to previous

visual world research on quantifiers embedded in referential phrases, our items had more complex stimuli. Interpreting a description like, "the girl that has all/some of the stripy squares" involves composing the meanings of a quantifier and modifier with the noun, rather than simply quantifier and noun. There were four identical agents in the display, two of which are still potential targets during the determiner window. Since the positioning of these two targets was randomly allocated across trials, participants who attempt to visually anticipate the target would be required to search around the whole display more during the instruction. Given the complexity of the items and the difficulty in anticipation in the determiner region, it may be that some participants were discouraged from attempting to compose determiners, *some* and *all* with the modifier prior to the noun and rather opted to exploit the modifier-noun composition – which completely disambiguates the referring expressions (i.e. only one girl has "stripy squares"). To address this issue, Experiment 3 uses less complex stimuli. It is designed so that there is only one target from the onset of the determiner region. It makes the modifier non-informative (a genitive phrase that can apply to all images in the display) so that the determiners become the focus of any anticipation by participants. In addition, in order to encourage anticipation and discourage participants waiting to hear the disambiguating noun, we allowed participants to control the pace of experiment such that at any point of an ongoing trial, as long as participants clicked on the image, the next trial began.

Experiment 3

Method

Participants
Thirty-six participants were recruited from our university campus via an online psychological subject pool. All participants speak English as a native language. They have uncorrected or corrected to normal vision.

Procedure
There were six practice trials in the beginning. Each practice trial began with a display depicting a character in the centre of the display who was about to distribute four sets of objects to two boys and two girls. Participants heard a background story describing the situation, for example, "This is Susan. She gives out fruits to children every day. Here is what she has on Monday. She has apples, pears, bananas, and oranges. She always brings more than enough. The leftover fruits are put in the middle". On the next display, the objects were
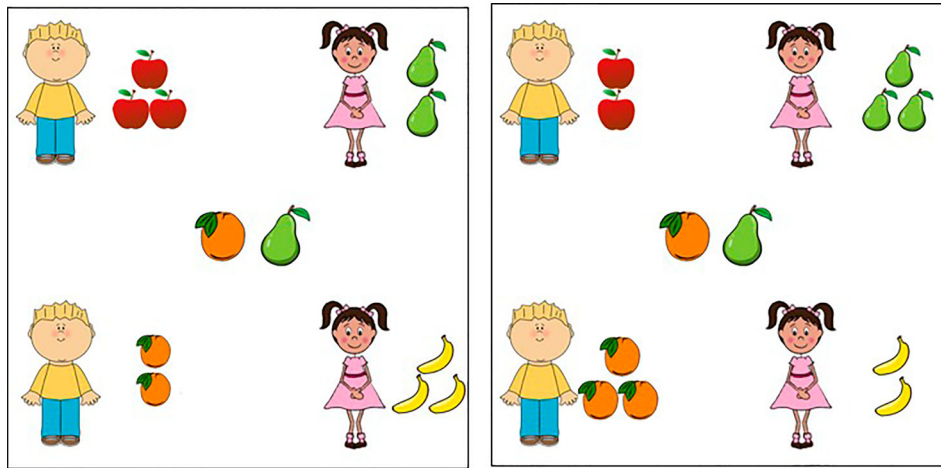
Figure 12. Example critical displays for Experiment 3. The left image (big *all*/ small *some*) can be paired with "Click on the boy that has all/three of Susan's apples" or "Click on the girl that has some/two of Susan's pears". The right image (small *all*/ big *some*) can be paired with "Click on the boy that has all/two of Susan's apples" or "Click on the girl that has some/three of Susan's pears".

distributed to boys and girls with the residue set in the centre. One second after the display onset, participants were given an auditory instruction, for example, "Click on the girl that has some of Susan's apples". Participants' task was to click on the image according to the instruction. The experimental script was set to jump to the next trial after participants clicked on the image. These six practice trials familiarised participants with the three characters (Susan, Amy, Michael) to be used in the experiment and the types of objects each character brings (fruits, stationery, kitchenware respectively).

After the practice session, we ensured that participants understood the story, instruction, display and procedure. Then the experiment began. The procedure was identical to the practice session except for one difference: On each trial, the background story and the starting display were not presented again. Participants were presented with the experimental display directly (see Figure 12). There were 48 trials, divided into 36 critical trials and 12 fillers. A randomised order of presentation of the items was created for each participant. The experiment was conducted using E-Prime software and a Tobii TX300 eye-tracker. Fixations were sampled every 17 ms. Calibrations were performed in the same way as in Experiment 2. For each trial, eye movements were recorded from the onset of the display to the point when the click occurred. The whole experiment lasted approximately 15 min.

Materials

The experiment employed three by two within-subject design. The two independent variables were *Determiner* (All, Some, Number) and *Target size* (Big, Small), which generated six experimental conditions: big *all*,

small *all*, big *some*, small *some*, big *number* (i.e. three), small *number* (i.e. two). The auditory instructions were of the form "Click on the [gender] that has [Det] of [name's] [object]". [gender] was either *boy* or *girl*, [Det] was one of *some*, *all*, *two*, *three*, [name's] was one of *Susan's*, *Amy's*, *Michael's*. Thirty-six experimental displays were constructed and paired with an audio instruction containing one of the determiners (e.g. "Click on the girl that has some of Susan's apples"). The experimental display contained four agents, two boys and two girls. As in Huang and Snedeker (2009), we arranged the display so that vertically adjacent agents were in the same gender and the horizontally adjacent characters were not. This means that participants could expect to locate boys and girls in the same locations on each trial. Four sets of objects were distributed among the agents. For two agents in the same gender, there was always one agent that had a total set of one kind of object and the other one had a proper subset. The residues of the two partitioned sets remained in the centre. In terms of set sizes, as in Experiment 2a, two agents always had a set of three objects and another two had a set of two objects. We counterbalanced the target set size for *all* and *some*. Figure 12 (left) can be used on a small-set *some* or big-set *all* trial and Figure 12 (right) can be used on a big-set *some* or small-set *all* trial. We defined two key time windows: the determiner window (from the determiner onset to "of" offset, e.g. during "some of") and the name window (from the name onset to "'s" offset, e.g. during "Susan's"). In both time windows, the target was the character of the description, the competitor was the character in the same gender.

Again three lists were created. Each list contained 36 experimental items, 12 items per determiner. In addition, each list contained 12 fillers. Fillers were similar to experimental items but contained different determiners (*One*, *Four*) in the instruction. The audio instructions were cross-spliced and adjusted as in Experiments 2a, b. The average duration for the determiner window was 708 ms (all of: 709 ms, some of: 711 ms, three of: 713 ms, two of: 700 ms), the average duration for the name window was 550 ms (Susan's: 551 ms, Michael's: 547 ms, Amy's: 552 ms). Each gender (boy/girl) was referred to an equal number of times within each condition. The scenario, the object and the location of the target were counterbalanced within each condition. All pictures of an agent with a set of objects measure 336*315 pixels. Picture of items in the middle measure 168*210. The screen resolution is 1680*1050 pixels.

Data treatment

We excluded one participant whose accuracy rate is lower than 2 standard deviations from the mean accuracy ($M = 35.13$, SD = 1.5). 1.7% of the trials were excluded because participants clicked on the wrong target. Twelve per cent of the trials were excluded due to track loss. Again, we first analysed participants' eye movements to target regions. Then we investigated participants' eye movements to the residue set.

Analyses of eye movements towards target region

Figure 13 visualised how target preference developed over time by determiner type and by determiner type and target size. We constructed separate linear mixed-effects models for each time window predicting target preference scores from fixed effects of *Determiner* (all, some or number), *Target size* (small or big), *Time* and their interactions, including maximal random effects structure supported by the data.

Determiner window

There was a significant effect of determiner type ($\chi^2(2) = 15.92$, $p = .003$). Post hoc analyses revealed a greater target preference in the *number* condition compared to the *all* and *some* conditions (all: $\beta = -0.90$, SE = 0.19, $t = -4.79$; some: $\beta = -0.94$, SE = 0.21, $t = -4.56$). Between *all* and *some*, the average target preference did not differ ($t = -0.17$, ns). There was also a significant effect of time ($\beta = 2.13$, SE = 0.41, $t = 5.16$), and a significant interaction of determiner type and time ($\chi^2(2) = 26.16$, $p < .001$). Post hoc analysis revealed that an increased target preference over time was only present in the

number condition ($\beta = 4.80$, SE = 0.69, $t = 6.94$), but not in the *all* or *some* condition (ts < 2). There was no difference on the linear slope over time between *all* and *some* ($t = -0.94$, ns).

We found a main effect of target size ($\beta = 0.42$, SE = 0.15, $t = 2.84$) and a significant interaction between target size and time ($\beta = 1.32$, SE = 0.62, $t = 2.12$), revealing a greater target preference with a steeper linear increase in the big-set target condition compared to the small-set target condition. We also found a significant interaction between target size and determiner type ($\chi^2(2) = 7.02$, $p = .03$) and a significant three-way interaction between target size, determiner type and time ($\chi^2(2) = 7.55$, $p = .02$). Examining each level of determiner type, we found that the target preference was greater and increased faster when *all* was used with a larger set compared to when it was used with a smaller set (overall bias: $\beta = 1.1$, SE = 0.26, $t = 4.17$; slope: $\beta = 3.96$, SE = 1.36, $t = 2.91$). In contrast, no significant difference was found between the big-set *some* and small-set *some* conditions, and between the *three* and *two* conditions (all |t|s < 2). We also found, regardless of the target set size, the number conditions showed a greater target bias and a steeper linear increase compared to the *all* and *some* conditions.[18] All other comparisons were non-significant.

Name window

The significant main effect of determiner type continued ($\chi^2(2) = 46.47$, $p < .001$), and there was a significant interaction between determiner type and time ($\chi^2(2) = 6.44$, $p = 0.4$). Post hoc analyses revealed that the target preference in the *number* condition was greater compared to the *all* and *some* conditions (all: $\beta = -1.86$, SE = 0.18, $t = -10.19$; some: $\beta = -2.05$, SE = 0.27, $t = -7.59$), but the rate of increase of target preference was faster in the *all* and *some* conditions compared to the *number* condition (all: $\beta = 1.91$, SE = 0.58, $t = 3.3$; some: $\beta = 1.74$, SE = 0.62, $t = 2.79$). No significant difference was found between the *all* and *some* conditions.

In the name window, there was no significant effect of target size or interaction between target size and time (|t|s < 2). There was, however, a significant interaction between target size and determiner type ($\chi^2(2) = 11.1$, $p = .004$). Pairwise comparisons revealed an effect of target size in the *all* condition ($\beta = 1.15$, SE = 0.46, $t = 2.5$), with greater target preference in the big-set target condition compared to the small-set target condition, but no effect of target size was found in the *some* or *number* condition. Further analyses on each level of target size revealed that when the target was a larger set, the target preference was significantly greater in the *all* condition compared to the *some* condition ($\beta = -0.86$, SE = 0.38, $t = -2.26$);
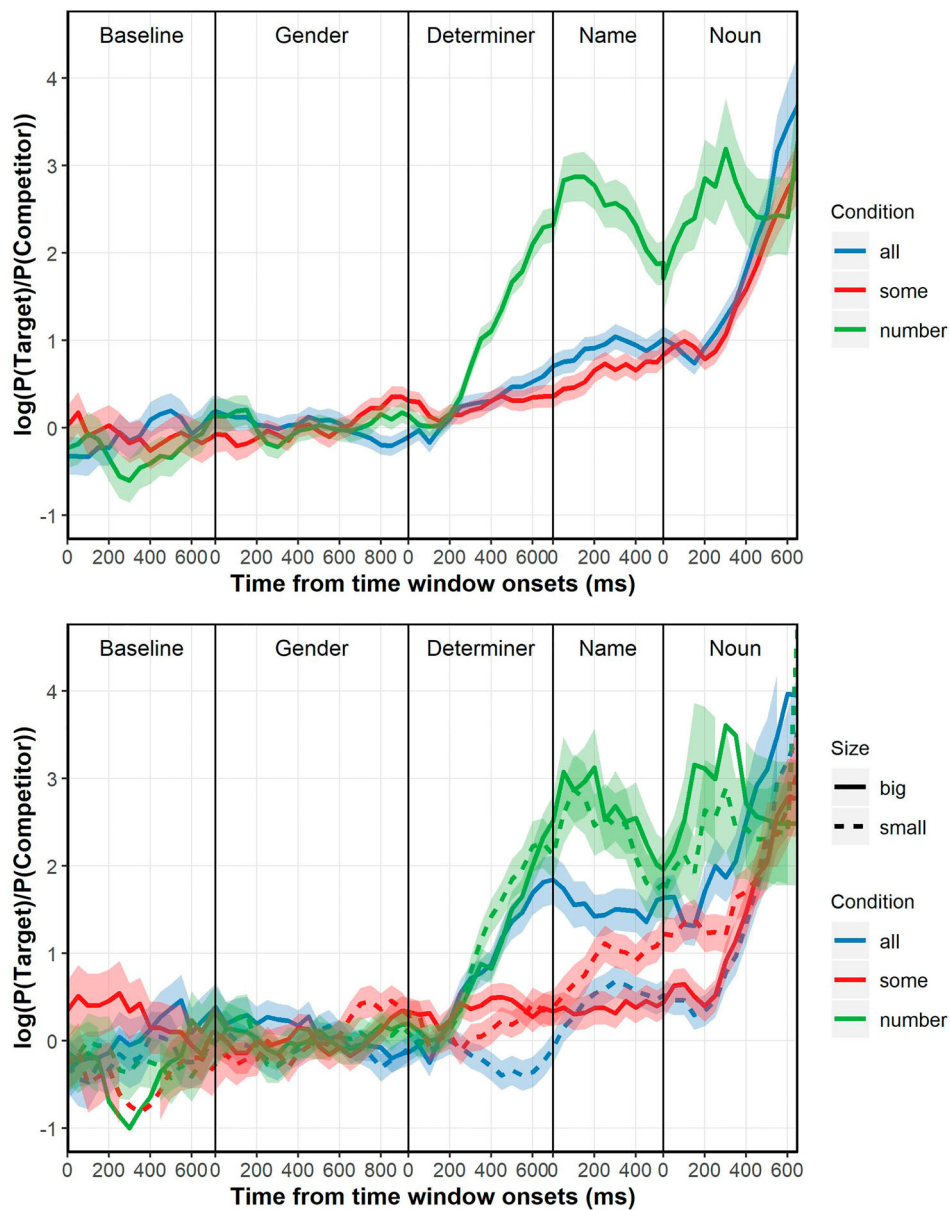
Figure 13. Target preference scores from the instruction onset to the instruction offset in Experiment 3. The top graph shows target preference scores by determiner type, and the bottom graph shows target preference scores by determiner type and target set size. Standard errors are represented by transparent ribbons.

whereas when the target was a smaller set, no difference in target preference was found between the *all* and *some* conditions ($t = 1.37$, n.s.). Similar to the previous window, regardless of the set size, target preference was always significantly greater in the *number* condition compared to both the *all* and *some* conditions.[19] There was no significant three-way interaction between target size, determiner type and time in this window.

Analyses of visual search to the residue set

Figure 14 shows the rise and fall in fixations to the residue in the *some*, *all* and *number* conditions over time. We constructed separate growth curve analyses for the determiner and name windows predicting fixation probabilities (empirical logit transformation) from fixed effects of *Determiner*, *Target size*, *Time* and their interactions. *Time* was represented by *Time 1* and *Time 2* to capture both the linear and quadratic change in fixations over time.

Determiner window

We found a significant interaction between *Time 1* and determiner type ($\chi^2(2) = 216.12$, $p < .001$). Post hoc analyses revealed that fixations on the residue set were decreased linearly only in the *number* condition ($\beta = -1.06$, $SE = 0.15$, $t = -7.17$), not in the *some* or *all*
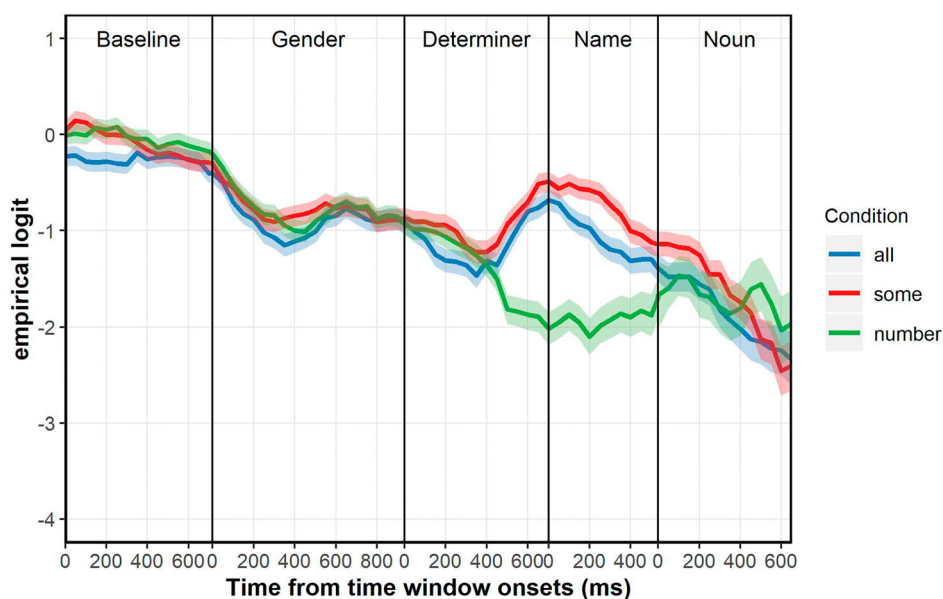
Figure 14. Fixation probabilities (empirical logit transformation) on the residue set by determiner type from the instruction onset to the instruction offset in Experiment 3. Standard errors are represented by transparent ribbons.

conditions. Also no difference in the linear slope was found between the *all* and *some* conditions ($t = -0.58$, n.s.). We also found a significant main effect of *Time 2* ($\beta = 0.31$, SE $= 0.10$, $t = 3.11$) and a significant interaction between *Time 2* and determiner type ($\chi^2(2) = 40.77$, $p < .001$). Post hoc analyses revealed significant positive quadratic coefficients for both *all* ($\beta = 0.51$, SE $= 0.13$, $t = 3.87$) and *some* ($\beta = 0.52$, SE $= 0.14$, $t = 3.66$) conditions, but not for the *number* condition ($t = -0.98$, ns). Other main effects and interactions were not significant.

Name window
We found a main effect of determiner type ($\chi^2(2) = 53.6$, $p < .001$), a main effect of *Time 1* ($\beta = -0.52$, SE $= 0.13$, $t = -4.15$) and a significant interaction between determiner type and *Time 1* ($\chi^2(2) = 29.96$, $p < .001$). Post hoc analyses revealed that fixations on the residue set were greater and decreased faster in the *all* and *some* conditions compared to the *number* condition (*all*: $\beta = 0.70$, SE $= 0.10$, $t = 6.92$; $\beta = -0.71$, SE $= 0.12$, $t = -5.63$; *some*: $\beta = 0.94$, SE $= 0.10$, $t = 9.18$; $\beta = -0.55$, SE $= 0.12$, $t = 4.40$). Between *all* and *some*, the average fixation was greater in the *some* condition ($\beta = 0.25$, SE $= 0.10$, $t = -2.46$), but there was no difference in the linear decrease between the *all* and *some* conditions. Neither the effect of *Time 2* nor the interaction between *Time 2* and determiner type was significant.

As in the previous experiments, we found no effect of set size on the overall bias. However, we found an effect of set size on the way bias formed over time in the name window where there was a significant interaction between target size and *Time 1* ($\beta = 0.21$, SE $= 0.10$, $t = $ 2.06) and a significant three-way interaction between determiner type, target size and *Time 1* ($\chi^2(2) = 15.97$, $p < .001$). Further analyses revealed that only in the *some* condition, there was a steeper decrease in the big-set target condition compared to the small-set target condition ($\beta = -0.63$, SE $= 0.24$, $t = -2.63$). Other main effects and interactions were not significant. We attribute this effect to the fact that residue-set search behaviour is completed earlier for *some* than for *all* in big-set trials since in *all* trials, the low-level association drives prolonged initial search of the target region.

Discussion

In Experiment 3, we broadly replicated the findings from Experiment 2. In the target region analyses, we found, in both time windows, a greater target bias with steeper linear increase in the *number* condition compared to the *all* and the *some* condition, and there was no difference in looking pattern between *all* and *some*. Concerning the influence of set size on the target bias, we found, the target bias was greater when *all* was used with the larger set compared to when it was used with the smaller set. We also found a greater target bias in the *all* condition compared to the *some* condition when both used with a larger set. Although, we did not detect the reverse pattern, found in Experiment 2a, of a greater bias in the *some* condition compared to the *all* condition when both used with a smaller set, these results confirm our hypothesis that a low-level association between *all* and larger sets has affected results in previous visual world studies.

In the analyses of visual search to the residue set, we found in the determiner window, there was an effect of *Determiner* on both the linear and quadratic terms of *Time*. On the linear term of *Time*, similar to Experiment 2a, only the *number* condition revealed a decrease in fixations towards the residue set; On the quadratic term of *Time*, unlike Experiment 2a, the *all* and the *some* condition revealed a U-shape parabolic change in fixation probability on the residue set. Since we find this searching behaviour for both *some* and *all* in contrast to *numbers*, the results confirm the predictions of the fast pragmatic view and disconfirm the predictions of the slow pragmatic view. The latter conclusion is based on the assumption that if *some* has only its literal meaning (*some and possibly all*) then participants need to wait until the noun for sufficient information to identify the target. In that case, visual search in *some* trials would not be directed at the residue set.

In the modifier window, the overall bias towards the residue set had become greater in the *all* and *some* conditions compared to the *number* condition. We also found greater bias in *some* condition than *all*. Again, this confirms the fast pragmatic view and tends to disconfirm the slow pragmatic view, which predicts that visual search of the residue set should be greater for *all*.

In contrast to the surprising finding of Experiment 2a, we did not find a positive coefficient of *Time 2* in the *number* condition. We suspect that no "return" effect might be due to the relatively short duration of the determiner window in Experiment 3. That is, in this experiment the determiner window (i.e. *all of/some of/ two of/ three of*) did not contain the definite article "the". This result also confirmed that the "return" effect in *numbers* found in Experiment 2a might largely due to the noise caused by specific design issues.

Finally, to determine whether the target referent was identified before the noun onset, we performed one-sample *t*-test to compare target proportions to chance (50%) over the combined window (from the determiner window onset to the name window offset). Results showed that the target proportion was significantly above chance for *all*, *some* and *numbers* (all: $t_1(34) = 3.26$, $p = .003$; $t_2(35) = 5.65$, $p < .001$; some: $t_1(34) = 4.81$, $p < .001$; $t_2(35) = 5.29$, $p < .001$; numbers: $t_1(34) = 12.35$, $p < .001$; $t_2(35) = 19.40$, $p < .001$). Thus, we found a significant bias to target in all three conditions before the disambiguating noun.

General discussion

This paper explored factors that could partly account for the mixed results in the timecourse of access and integration of pragmatically enriched *some*. In previous visual-world studies that included items with number terms as well as *some* and *all*, Huang and Snedeker (2009, 2011) reported a delay in pragmatic *some* compared to *all*, whereas Degen and Tanenhaus (2016) found a delay only when the target was a larger set in the display but not when it was a smaller set. We hypothesised that people develop expectations about the target based on low-level associations between relative set size and *all*. In addition, we noted that factors such as Maximise Presupposition could drive attention away from targets with sets of two objects. In Experiments 1a,b, we tested our hypotheses about these factors and demonstrated that, independently of processes which establish the asserted truth conditions, some combination of low-level associations and Maximise Presupposition influence participants' responses. In Experiments 2a,b, we explored our hypothesis about previous results in a visual world design and showed such prior associations influenced the target bias formation for the *all* condition. These findings render the interpretation of previous visual world data problematic. When set size is not controlled, as in Huang and Snedeker (2009), the delay in target identification for *some* relative to *all* could be partly due to these factors rather than the slow pragmatic calculation.

To address the issue that bias formations were affected by low-level associations, in Experiments 2 and 3, we fully counterbalanced the target set size for the *all* and *some* referents and explicitly modelled the target size in the analyses. We found that, when set size was controlled, the timecourse of looks to the target based on enriched-*some* is not different from that for *all*.

In Experiments 2 and 3, we introduced a novel indicator to measure the timecourse of scalar processing. This was looks to the region where the residues of the partitioned sets are located. Critically, visual search to the residue set allowed us to test different predictions made by the fast-pragmatic account and the slow-pragmatic account. Using visual search to the residue set in number trials as a baseline, after the determiner onset, both fast- and slow-pragmatic accounts predict a difference between *all* and *numbers* such that greater visual bias forms to the residue set for *all* and search patterns should be reflected in a U-shaped pattern in timecourse data. As for *some* trials, the fast-pragmatic account predicts the same pattern as for *all trials*. By contrast, the slow-pragmatic account predicts less visual bias in *some* than *all*, and no U-shaped search pattern. Our results show positive evidence for the fast-pragmatic account and negative evidence for the slow-pragmatic account. In particular, for all three experiments, we found a greater bias to the residue set in *some* and *all*

compared to *number* trials. In addition, changes of directions in fixation data, which is an initial decrease followed by an increase in residue set fixations in *some* and *all* trials provide further evidence that the enriched-*some* is accessed and integrated rapidly.

Therefore, across three visual-world studies, the target preference and residue set analyses provide consistent and converging evidence that an enriched interpretation of *some* is accessed in the same timecourse as literal interpretations of *all*. In addition, our residue set analyses provide evidence against the idea that *some* is initially analysed as "some and possibly all".

Compared to several previous studies that examined the timecourse of processing *some* and *all* vs. number, we find a consistent advantage in number trials over *some* and *all* trials. As explained above, other things equal, this is a pattern we should expect because the search behaviour required for numbers is simpler than that for *all* and *some*. As we noted in our discussion of Experiment 2a, a comparison between big *all* trials (set size = 3) and numbers showed no difference and this is the comparison made in Huang and Snedeker (2009), which found the same pattern. We attributed the "improved" performance for big *all* trials to a boost obtained by other factors.[20]

In this paper, we have established that we get a better account of previous results of timecourse studies that included numbers by factoring in a low-level association between *all* and set size and Maximise Presupposition effects. Our research has focused on studies that include number items because these have been the studies which have previously shown a difference between *some* and *all*. According to our proposals, studies which do not involve numbers and do not counterbalance set size should be equally affected by these low-level effects. However, some such studies have not shown differences between *all* and *some* conditions (e.g. Breheny et al., 2013; Grodner et al., 2010). One explanation for this result follows Huang and Snedeker (2018) in assuming that when numbers are absent, participants are more likely to engage in precoding than when they are absent.

To conclude, the debate about the timecourse of the availability of an enriched meaning of *some* compared to *all* has recently been focused on explaining discrepant results. In this paper, we have argued that when set-size is not counterbalanced, a low-level association between *all* and larger set has been responsible for the apparent advantage for *all* over *some*. In addition, in studies where a set size of 2 has been used for *some* trials but not *all* trials, the effect has been exacerbated by a dis-preference for *all* to be associated with sets of 2 (Maximise Presupposition). Once these factors are controlled for, we find no difference in target bias. In addition, our novel design which allows us to measure attention to the residue set has provided clear evidence that the enriched meaning of *some* is accessed in the same timecourse as the meanings for *all* and numbers. At the same time, U-shaped search pattern for *some* and *all* trials provides disconfirming evidence for the alternative, slow-pragmatic account.

Notes

1. Specifically, a strong case has been made that scalar implicature is mediated by linguistically represented operators (see Chierchia, Fox, & Spector, 2012, a.o.). According to that account, both unenriched and enriched *some* are derived via the grammar. However, even the grammatical account recognises an enhanced role for contextual inference in the enriched case, in terms of computing and selecting alternatives according to relevance (see Breheny, 2019 for discussion). Recent Bayesian probabilistic accounts (see Bergen, Levy, & Goodman, 2016; Potts, Lassiter, Levy, & Frank, 2016) likewise imply an enhanced computation involving alternatives for the enriched case.
2. Or at least the two should not differ before the final noun onset.
3. However, the possibility of pre-coding could not be completely ruled out in the number absent study for two reasons: first, although "all" and "some" were used with, respectively subsets and total sets each for one quarter of the time, pre-coding the set types can still be beneficial in these decoy cases as it facilitates the rejection response. Second, if participants labelled the subset with "some", we would expect that they were more likely to provide the "false" response (i.e. click on the central button) to pragmatically infelicitous statements compared to the same stimuli in the number present study. Degen and Tanenhaus found that the rate of "false" responses in number absent study was indeed higher than in number present study. Even though the naturalness hypothesis can in part explain these findings, these results are also compatible with the view that participants pre-coded the subset with some.
4. For instance, (1) is infelicitous because other things equal, the alternative (2) carries a uniqueness presupposition.
    (1) # A sun is shining.
    (2) The sun is shining.
5. Similarly, in Degen and Tanenhaus (2016), "You got all of …" was paired with a set of two gumballs in one-fourth of the critical *all* trials.
6. We note that the results of Experiments 1a,b could be explained by an alternative account: that Maximise Presupposition alone is responsible for the effect with *all* in Experiment 1a and there is a low-level preference for *all* to be used with numbers 4 or more. We note also that if correct, the account equally calls into question the results of previous visual world studies and predicts the patterns of results we obtained in the visual world

studies reported below. We thank an anonymous reviewer for pointing out this alternative.

7. We note here that because Experiments 1a,b did not contain number items and because the objects in the sets are different, we are not, strictly speaking, basing the predictions for Experiments 2a,b directly on the results of Experiments 1a,b. However, Experiments 1a,b provides further evidence for factors we hypothesised to account for results of previous studies. Experiments 2a,b provide a further, systematic test of that hypothesis using visual world methods.

8. We note that participants can determine what are the total set sizes prior to the onset of the linguistic stimuli. If a participant does that, then they do not need to consult the residue set during the utterance. However, to the extent that participants do not memorise the visual state of affairs, or need to double check, then it is the compositional semantics of *all* and enriched *some* vs. numbers which will guide their search behaviour.

9. Each audio instruction was first recorded individually. Then we spliced determiner, modifier and shape words into an instruction schema created from a recording of, "Click on the girl that has most of the orange oblongs", which provides no advantage to any condition in terms of co-articulation information prior to critical words.

10. We excluded trials with greater than 25% track loss (when the eye-tracker captured participants' gaze location with very low validity or when participants looked at non-AOI).

11. Note that given our design, we had a choice of competitor for the target in the modifier window, for example, in the *some*-stripy trials one competitor could be the *all* girl with the same pattern (stripy) and the other could some girl with a different pattern (dotted). We accept that, having processed the determiner, participant bias is already moving away from determiner-inconsistent targets (away from all girls) but we know from previous work that while the composition of quantifier with nominal affects bias in an incremental fashion, it does not fully constrain gaze (Altmann & Kamide, 2007). At the same time, we know that bottom up linguistic processing (here of "stripy") can attract attention in a way that is only modulated by biases from previous processes (see e.g. Barr, 2008a). Thus we expect that our DV here should reveal the relative strength of bias resulting from previous processes integrating visual context, linguistic and pragmatic information.

12. *Determiner* was then coded using treatment coding, with *all* or *number* condition as the reference level (depending on the pairwise comparison).

13. $\ln\left(\frac{y + 0.5}{n - y + 0.5}\right)$ where *n* refers to the total number of looks to AOIs in the visual display, and *y* refers to the number of looks to the residue set area.

14. When the target was a larger set, *all*: $\beta = -0.52$, SE = 0.23, $t = -2.23$; *some*: $\beta = -1.15$, SE = 0.24, $t = -4.71$; When the target was a smaller set, *all*: $\beta = -2.22$, SE = 0.29, $t = -7.75$; *some*: $\beta = -1.30$, $SE = 0.28$, $t = -4.71$.

15. We thank an anonymous reviewer for suggesting that we consider this alternative account.

16. When the target was a larger set, *all*: $\beta = -1.10$, SE = 0.26, $t = -4.16$; *some*: $\beta = -2.05$, SE = 0.28, $t = -7.28$; When the target was a smaller set, *all*: $\beta = -1.56$, SE = 0.29, $t = -5.38$; *some*: $\beta = -0.86$, SE = 0.30, $t = -2.87$.

17. Other studies, such as Degen and Tanenhaus (2016) do include residue sets but attention to these was not analysed.

18. When the target size was big, *three* vs. *all*: $\beta = -0.40$, SE = 0.16, $t = -2.49$ (overall), $\beta = -2.6$, SE = 0.86, $t = -3.03$ (slope); *three* vs. *some*: $\beta = -0.80$, SE = 0.21, $t = -3.78$ (overall), $\beta = -3.27$, SE = 1.17, $t = -2.80$ (slope). When the target size was small, *two* vs. *all*: $\beta = -1.34$, SE = 0.28, $t = -4.84$ (overall), $\beta = -5.95$, SE = 0.91, $t = -6.54$ (slope); *two* vs. *some*: $\beta = -1.31$, SE = 0.25, $t = -5.16$ (overall), $\beta = -4.91$, SE = 0.89, $t = -5.48$ (slope).

19. When the target size was big, *all*: $\beta = -1.66$, SE = 0.23, $t = -7.02$; *some*: $\beta = -2.55$, SE = 0.36, $t = -7.13$; When the target size was small, *all*: $\beta = -2.15$, SE = 0.33, $t = -6.59$; *some*: $\beta = -1.76$, SE = 0.27, $t = -6.57$.

20. We note here that Degen and Tanenhaus (2016) results appear to show a distinct advantage for numbers over the other determiners when set size was small, although the results of any relevant comparison were not reported.

Disclosure statement

Funding

References

Altmann, G., & Kamide, Y. (2009). Discourse-mediation of the mapping between language and the visual world: Eye movements and mental representation. *Cognition*, *111*(1), 55–71.

Altmann, G. T., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, *57*(4), 502–518.

Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, *30*(3), 191–138.

Arai, M., van Gompel, R. P., & Scheepers, C. (2007). Priming ditransitive structures in comprehension. *Cognitive Psychology*, *54*, 218–250.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Barr, D. J. (2008a). Pragmatic expectations and linguistic evidence: Listeners anticipate but do not integrate common ground. *Cognition*, *109*(1), 18–40.

Barr, D. J. (2008b). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*(4), 457–474.

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*(3), 255–278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bergen, L., Levy, R., & Goodman, N. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, *9*, 1–83.

Boersma, P., & Weenink, D. (2017). Praat: Doing phonetics by computer [Computer Program]. Version 5.1.40. Retrieved from http://www.praat.org/

Breheny, R. (2019). Scalar implicature. In N. Katsos & C. Cummins (Eds.), *Oxford handbook of experimental semantics and pragmatics* (pp. 39–62). Oxford: Oxford University Press.

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

Chierchia, G., Fox, D., & Spector, B. (2012). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In P. Portner, C. Maienborn, & K. von Heusinger (Eds.), *An international handbook of natural language meaning* (pp. 2297–2332). Berlin: Mouton de Gruyter.

Degen, J., & Tanenhaus, M. K. (2011). Making inferences: The case of scalar implicature processing. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 3299–3304). Austin, TX: Cognitive Science Society.

Degen, J., & Tanenhaus, M. K. (2015). Processing scalar implicature: A constraint-based approach. *Cognitive Science*, *39*(4), 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*(1), 172–201.

Dehaene, S. (1997). *The number sense: How the mind creates mathematics, revised and updated edition*. New York, NY: Oxford University Press.

Grice, H. P. (1967). Logic and conversation. *Syntax and Semantics*, *3*, 43–58.

Grice, H. P. (1989). *Studies in the way of words*. Cambridge: Harvard University Press.

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

Hallett, P. E. (1986). Eye movements. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance I: Sensory processes and perception* (pp. 10–102). New York, NY: Wiley.

Heim, I. (1991). Artikel und definitheit. In A. von Stechow & D. Wunderlich (Eds.), *Semantik: ein internationales Handbuch der* (pp. 487–535). Berlin: de Gruyter.

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

Huang, Y. T., & Snedeker, J. (2011). *Logic and conversation* revisited: Evidence for a division between semantic and pragmatic content in real-time language comprehension. *Language and Cognitive Processes*, *26*(8), 1161–1172.

Huang, Y. T., & Snedeker, J. (2018). Some inferences still take time: Prosody, predictability, and the speed of scalar implicatures. *Cognitive Psychology*, *102*, 105–126.

Kurumada, C., Brown, M., Bibyk, S., Pontillo, D. F., & Tanenhaus, M. K. (2014). Is it or isn't it: Listeners make rapid use of prosody to infer speaker meanings. *Cognition*, *133*(2), 335–342.

Mirman, D. (2017). *Growth curve analysis and visualization using R*. Boca Raton, FL: CRC Press.

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*(4), 475–494.

Politzer-Ahles, S., & Fiorentino, R. (2013). The realization of scalar inferences: Context sensitivity without processing cost. *PloS One*, *8*(5), e63943.

Potts, C., Lassiter, D., Levy, R., & Frank, M. C. (2016). Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. *Journal of Semantics*, *33*(4), 755–802.

Rayner, K., & Pollatsek, A. (1989). *The psychology of reading* (pp. 113–187). Englewood Cliffs, NJ: Prentice-Hall.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Sauerland, U. (2008). Implicated presuppositions. In A. Steube (Ed.), *The discourse potential of underspecified structures* (pp. 581–600). Berlin: Mouton de Gruyter.

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, *32*(1), 3–23.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109–147.

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*(5217), 1632–1634.

van Tiel, B. (2014). Embedded scalars and typicality. *Journal of Semantics*, *31*(2), 147–177.