Implicature priming, inverse preference, and context adaptation

Paul Marty – paul.marty@um.edu.mt

Institute of Linguistics & Language Technology
Faculty of Media & Knowledge Sciences, Room 602
L-Università ta' Malta
Msd 2080, L-Imsida, Malta

Jacopo Romoli – jacopo.romoli@hhu.de

Institut für Sprache und Information
Gebäude 23.21, Ebene 04, Raum 00.77
Heinrich-Heine-Universität Düsseldorf
40225, Düsseldorf, Germany

Yasu Sudo – y.sudo@ucl.ac.uk

Division of Psychology & Language Sciences
Chandler House, Room 115H
University College London
WC1N 1PF, London, UK

Richard Breheny – r.breheny@ucl.ac.uk

Division of Psychology & Language Sciences
Chandler House, Room 108
University College London
WC1N 1PF, London, UK

Abstract

Previous studies found that scalar implicatures (SIs) arise more often after Strong priming trials that force the derivation of an SI than after Weak priming trials that force their inhibition. Previous proposals assume that priming is of a sub-mechanism of SI calculation that references alternatives. The present study tests this proposal by comparing the effects of Strong and Weak primes to baseline conditions that involve no priming. Our results establish that implicature priming effects are *inverse preference effects* and they show that, when the preferred interpretation in the baseline conditions is the one with the SI, the observed priming effects are in fact inhibitory effects driven by the Weak primes. Capitalising on the notion of *context adaptation*, we propose a novel account of these effects on which implicature priming is explained in terms of rapid and incremental adaptation of language users' expectations about the type of context they are in, rather than in terms of boosting the activation level of alternatives.

Keywords: scalar implicatures, priming effects, inverse preference, context adaptation

1. Introduction

Over the past 20 years, *scalar implicatures (SIs)* have garnered a considerable amount of attention in Experimental Pragmatics. A prototypical example of SI comes from the word *some* in English. Concretely, there are two ways of understanding a sentence such as *Some of the symbols are circles*. On its *strong reading*, the sentence means that some but not all of the symbols are circles, while on its *weak reading*, it is read as simply asserting the existence of circles, and is neutral with respect to whether all of the symbols are circles or not. A number of previous experimental studies have established the availability of both of these readings across different experimental tasks (see Bott & Noveck 2004; Katsos & Cummins 2010; Chemla & Singh 2014; Noveck 2018 and references therein).

Since the seminal work by Grice (1989), it is considered that the weak reading is the literal interpretation of the sentence, and the strong reading involves the computation of an additional inference—namely, the SI—which amounts to the negation of the version of the sentence where *some* is replaced with *all*. More generally, SIs can be characterised as the negations of such *alternatives*, which are related sentences whose negations are consistent with the literal meaning of what is asserted. It is currently actively debated in the theoretical literature how exactly such inferences about the negations of alternatives are computed, especially regarding the question of the extent to which this inferential process is to be seen as rooted in contextualised pragmatic reasoning about the speaker's communicative intention and a specific conversational background (Sauerland, 2004; Chierchia, Fox & Spector, 2012; Geurts, 2010; Bergen, Levy & Goodman, 2016). Since this theoretical question is orthogonal to the main interest of the present paper, we will largely remain neutral about it.

For the past two decades, there has been a surge of experimental studies on SIs, aiming at shedding light on different theoretical questions. Among the many experimental techniques that have been employed to investigate different aspects of SI, this paper focuses on the *priming paradigm* (Bott & Chemla, 2016; Rees & Bott, 2018; Waldon & Degen, 2020; Meyer & Feiman, 2021). The previous studies on implicature priming cited here commonly observe the following general pattern: SIs arise more often after *strong primes*—priming trials that force the strong reading and thus the computation of an SI—than after *weak primes*—priming trials that force the weak reading, which involves no SI. In order to explain this observation, Bott & Chemla (2016) put forward an account in terms of priming of (certain aspects of) the computational mechanism used to generate SIs (which is adopted by later studies such as Rees & Bott 2018; Meyer & Feiman 2021). More specifically, as explained above, the computation of an SI involves (i) referencing an alternative and (ii) negating it. Potentially, both of these aspects could be primed, but Bott & Chemla (2016) argue that the former aspect about alternatives gives rise to a particularly robust priming effect by increasing the salience level of the alternative referenced.[1] Specifically, a strong prime forces an SI to be computed, which by assumption references an alternative (e.g., the sentence with *all*, in the

[1] Bott & Chemla (2016) claim that priming of the mechanism that is involved in the computation of SIs is also possible and explain part of their experimental results in terms of it. However, its effect size is expected to be very small. We will discuss their account in greater detail later in the paper.

case of the above example), and this alternative *all* is thereby made salient or activated, and hence more likely to be referenced again in a different trial following the priming trial.

According to such an activation-based account, the main priming effects are driven by the strong primes' boosting effects, which increase the activation level of the relevant alternatives. Thus the weak primes, which do not force SIs to be computed and hence do not require alternatives to be referenced during their interpretation, should not have large effects. We tested this prediction by comparing the effects of strong and weak primes with a baseline condition. Our baseline condition was administered before participants encountered any priming trial and so involved no priming whatsoever. Our results show that for, certain scalar expressions, the strong primes indeed give rise to boosting effects. However, for others, it is in fact the weak primes that drive the main effect by inhibiting SIs. Overall, the observed priming effects are best characterised as *inverse preference effects* in the sense that the less dominant reading in the baseline condition gives rise to sizable priming effects whereas the dominant reading in the baseline condition hardly has priming effects, if any.

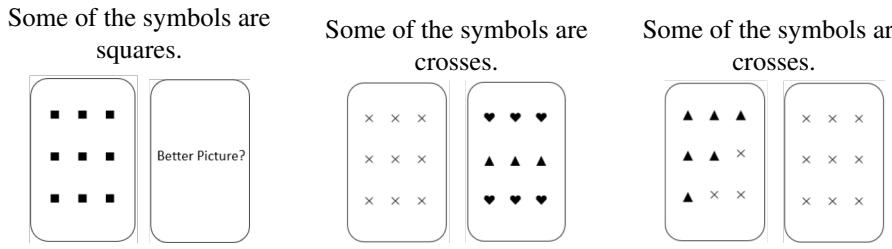
While inverse preference in the general sense is compatible with accounts that are framed in terms of priming the SI mechanisms, as Bott & Chemla (2016) discuss, we claim that the activation-based account does not provide a satisfactory explanation of the full set of data. In particular, it fails to explain the robust priming effects of weak primes in our results.

We propose an alternative explanation that is based on what we call *context adaptation*, which builds upon the adaptation-based view of priming put forward by Fine, Jaeger, Farmer & Ting (2013) and Jaeger & Snider (2013), among others, for inverse preference effects previously observed for syntactic priming (see also Waldon & Degen 2020, who propose a related idea for implicature priming). According to these authors, what gives rise to inverse preference effects is implicit learning and rapid online adaptation of probabilistic expectations about how often words and constructions are used in the current conversational context, and more generally about what kind of conversational context one is likely to be in. We argue that the inverse preference effects for implicature priming are amenable to a similar adaptation-based explanation. To give further support for our analysis based on context adaptation, we present *Hidden Markov Models* as computational models of context adaptation.

2. Implicature priming

Since our experiments are designed after the priming experiments reported in Bott & Chemla (2016), we first review the experimental paradigm in some detail. Bott & Chemla employed a picture selection task where each trial had two pictures, one overt and one hidden, the latter with a label 'Better Picture?', as in Figure 1a. These pictures were presented together with a sentence that can potentially have an SI, such as *Some of the symbols are squares*. Crucially, the overt picture is only compatible with the weak reading of this sentence, so that the participant would choose the overt picture only if they considered the weak reading to be an acceptable reading of the sentence.

Note that even when the participants choose the overt picture, it is still possible that they have accessed the strong reading, because they should choose the overt picture as



(a) Target (b) Weak Prime (c) Strong Prime

Figure 1: Example items illustrating the logic behind the SOME condition of Bott & Chemla (2016) (these are not the actual items they used).

long as they deem the weak reading acceptable and they might have noticed that the sentence is ambiguous. Therefore, the rate at which the overt picture is chosen should be understood as how often the weak reading is accepted, rather than the rate at which the strong reading was not accessed. Similarly, when the participants choose the hidden picture, that will indicate that they considered the weak reading to be unacceptable, but will not strictly speaking guarantee they accessed the strong reading. However, we take a hidden picture choice as a proxy measure for accessing the strong reading, assuming that the only reason to consider the weak reading inaccessible is due to the availability and acceptability of the strong reading.

Each target item of Bott & Chemla's experiments was preceded by two priming trials of the same kind. Priming trials also involved two pictures, but both overt. There are two types of priming trials, *weak primes* and *strong primes*. In a weak prime, only one of the two pictures makes the sentence true, and the other one renders the sentence clearly false. Crucially, however, the picture that makes the sentence true does so only under the weak reading. For example, for the sentence *Some of the symbols are crosses*, all the symbols on the card that makes the sentence true are crosses, while the other card clearly falsifies the sentence, i.e. none of the symbols are crosses. See Figure 1b, for example. Consequently, the participant is forced to choose the first picture, which in turn means they have to access the weak reading. In a strong prime, the sentence can be true with respect to both of the overt pictures. One of the pictures is the same as in the corresponding weak prime, and the sentence is only true if it is understood on its weak reading. Crucially, the second picture is one where the sentence is true on both weak and strong readings, e.g. for the sentence *Some of the symbols are crosses*, some but not all of the symbols in this picture are crosses. Since the participant was told that one and only one picture makes the sentence true, they are forced to understand the sentence under the strong reading and choose the picture that makes the strong reading true.

Using this paradigm, Bott & Chemla (2016) tested three types of scalar items: AD HOC, SOME, and NUMBER. An example sentence of each type is given in (1).[2]

[2]In addition to these three, Bott & Chemla (2016) tested another type of scalar item, PLURAL. It involved a plural noun phrase (e.g., *There are circles*) and the relevant inference was the plurality inference (e.g.,

(1) a. There is a dot. AD-HOC
b. Some of the symbols are clubs. SOME
c. There are four triangles. NUMBER

The relevant scalar implicature for these example sentences are the following.[3]

(2) a. ¬(There is a star). AD-HOC
b. ¬(All of the symbols are clubs). SOME
c. ¬(There are five triangles). NUMBER

The results of Bott & Chemla's experiments indicate that, for all three scalar expressions, the hidden picture was chosen more often in target trials following strong primes than in those following weak primes. This difference between weak and strong primes has been replicated by later studies (Rees & Bott, 2018; Waldon & Degen, 2020; Meyer & Feiman, 2021). Furthermore, a difference in the same direction was observed even when the target trial involved a different scalar item than the priming trials ('cross-scale priming'), albeit the size of the difference was considerably smaller than when the target and priming trials involved the same scalar item ('within-scale priming') (see also Meyer & Feiman, 2021).

The fact that cross-scale priming was observed led Bott & Chemla to conclude that there is a common mechanism behind the relevant inferences of the three types of scalar items, and that certain aspects of it can be primed in the experimental paradigm under discussion. Recall that generally the computation of a given SI involves (i) referencing an alternative and (ii) negating it. Since different scales involve different alternatives, what is common across scales must be the mechanism that searches for and negates alternatives. Thus, Bott & Chemla interpreted the results of cross-scale priming as coming from priming involving the use of this common mechanism. They further hypothesise that the comparatively large effect size observed with within-scale priming is due to priming the use of a particular alternative. Specifically, they claim that a strong prime forces the participant to compute an SI, which requires the above two steps, (i) and (ii). The common computation mechanism for SIs gives rise to a priming effect to the same degree as in the case of cross-scale priming, but in addition, referencing an alternative boosts the salience level of that alternative, making it more active and thus more likely to be used when the same scalar item is encountered afterwards (see Rees & Bott 2018; Waldon & Degen 2020 for related claims). This way, Bott & Chemla account for the larger effect size of within-scale priming as a combined effect of two

that there is more than one circle), which can be seen as an SI that is derived from the singular alternative. However, their experimental results indicate that PLURAL behave differently from the other three types of expressions they tested, especially with respect to 'cross-scale priming' (see below). We will therefore not discuss PLURAL in this paper.

[3]Two caveats are in order. Firstly, one might hesitate to call AD HOC a *scalar item*, but it could actually be seen as involving a contextually determined scale (see Hirschberg, 1991). For instance, one can understand the scale for (1-a) as consisting of ⟨There is a dot, There is a dot and a star⟩. Secondly, following Bott & Chemla (2016), we tentatively assume for now that (2-c) is an SI of NUMBER, noting that this is potentially theoretically controversial. According to some theories (e.g., Geurts, 2006; Breheny, 2008), (2-c) is (or can be) part of the literal meaning (see Spector, 2013, for an overview of different theories). We will come back to this debate at the end of the paper.

types of priming.

However, we would like to point out that there is an important gap in our current understanding of implicature priming: While Bott & Chemla (2016), as well as other previous studies on implicature priming, claim that implicature generation can be primed based on the difference between weak and strong primes, it is not entirely clear if the effects are driven by strong primes boosting the rate of SI generation, or weak primes lowering it, or both[4] This question is of high theoretical relevance, because if it turns out that the effects are primarily driven by weak primes inhibiting SI computation and strong primes have almost no boosting effects, for example, Bott & Chemla's explanation will have to be reconsidered, as it relies on the assumption that strong primes have large boosting effects, because they increase the activation level of an alternative.

Waldon & Degen's study sheds some light on this question. Unlike the other studies mentioned above, their experiment included a BASELINE condition in which each target item was preceded by a math task, such as '4 + 5 =?', and the participants choose one of two cards with numbers that had the correct answer.[5] The idea is that unlike priming trials with linguistic material, such math tasks should be neutral with respect to implicature priming. The results of this experiment show that the rate of strong readings in the BASELINE condition is intermediate between those of the target items after strong primes and the target items after weak primes, which seems to give some credence to the idea that implicature priming involves both boosting effects of strong primes and inhibition effects of weak primes.

However, there is a potential complication with this interpretation of Waldon & Degen's results due to what we call *spillover effects*: Given that items in the BASELINE condition were interspersed with the rest of the trials in Waldon & Degen's experiment, we cannot exclude the possibility that priming effects of preceding priming trials were present in the target items of the BASELINE condition that followed them, at least in some cases, and consequently the results from the BASELINE condition might not be completely neutral with respect to priming effects. Certainly, it seems that priming effects generally do decay with time, although this aspect is less investigated for implicature priming, in comparison to syntactic priming. However, at the same time, certain cases of syntactic priming are known to keep having effects across many sentences, or sometimes even for multiple days (Bock & Griffin, 2000; Bock, Dell, Chang & Onishi, 2007; Branigan, Pickering, Steward & Mclean, 2000; Kaschak, 2007; Kaschak, Loney & Borreggine, 2006; Kaschak & Borreggine, 2008; Kaschak, Kutta & Jones, 2011). It is possible, therefore, that the items in the BASELINE condition of Waldon & Degen's experiment were not completely free of priming effects coming from preceding trials, and comparing this condition to the priming conditions might not provide a complete answer to question we are after.

[4]Bott & Chemla (2016) and Rees & Bott (2018) have raised this issue before us, but see below.

[5]The experiments reported in Bott & Chemla (2016) did not have baseline conditions, but in their discussion of inverse preference, they made use of the fact that PLURAL exhibited no significant cross-scale priming effects (cf. fn. 2), and took the results of the PLURAL cross-scale priming trials as a proxy baseline for SOME and NUMBER, and the results of the SOME and NUMBER priming trials as a proxy baseline for PLURAL. Since these priming trials were interspersed with all the other trials, however, our concern about 'spillover effects' discussed below applies to this case as well.

3. Experiment 1: Probing the Direction of Priming Effects

Our first experiment is aimed at determining the direction of priming effects using a more neutral baseline than Waldon & Degen (2020) (and Bott & Chemla 2016; see fn. 5). To achieve this, we adopted a two-block design: The first block of the experiment involves no priming whatsoever, meaning all the items are trials with covered pictures; Then priming trials were introduced in the second block. Block 2 was similar to Bott & Chemla's (2016) original design in all relevant regards. This way, we could obtain more pristine baseline rates from the first block, against which the priming effects observed in the second block would be compared.

3.1. Data availability

Stimuli, data and analysis code associated with Experiment 1 are available open access on the OSF at https://osf.io/263xf/.

3.2. Methods

3.2.1. Participants

56 self-reported native speakers of English participated in this experiment (36 female, average age 32.6 years). Participants were recruited online through Prolific (UK/US IP addresses; minimum 90% prior approval rating). Participants were paid £1.40, and average completion time was about 9 minutes. Participants gave written informed consent. Data were collected and stored in accordance with the provisions of Data Protection Act 2018, the UK's implementation of the General Data Protection Regulation. The experiment was approved by the Research Ethics Committee at UCL.

3.2.2. Materials

Each trial involved a sentence presented above two pictures (see Figure 2). Sentences in the experimental trials were constructed using one of the three frames in (3) (in common with Bott & Chemla (2016), Rees & Bott (2018) and Waldon & Degen (2020)). The [symbol] term was a noun denoting a symbol type from the following list: arrow, cross, circle, diamond, heart, square, star or triangle.

(3) a. There is a [symbol]. AD-HOC
b. Some of the symbols are [symbol]. SOME
c. There are four [symbol]. NUMBER

Pictures consisted of a rectangle containing either symbols, henceforth symbol cards, or the text 'Better Picture?', henceforth the covered card. Symbol cards could be false, strong or weak. Priming trials consisted of two symbol cards: Weak priming trials involved a weak and a false card, and Strong priming trials involved a weak card and a strong card. Control and Target trials consisted of the covered card and one symbol card: a strong card in the True control trials, a false card in the False control trials and a weak card in the target trials. Example control, priming and target trials used in Experiment 1 are given in Figure 2.

For *Ad-hoc* trials, weak cards contained two different symbols, one of which matched the [symbol] term in the accompanying sentence. Strong and false cards contained a

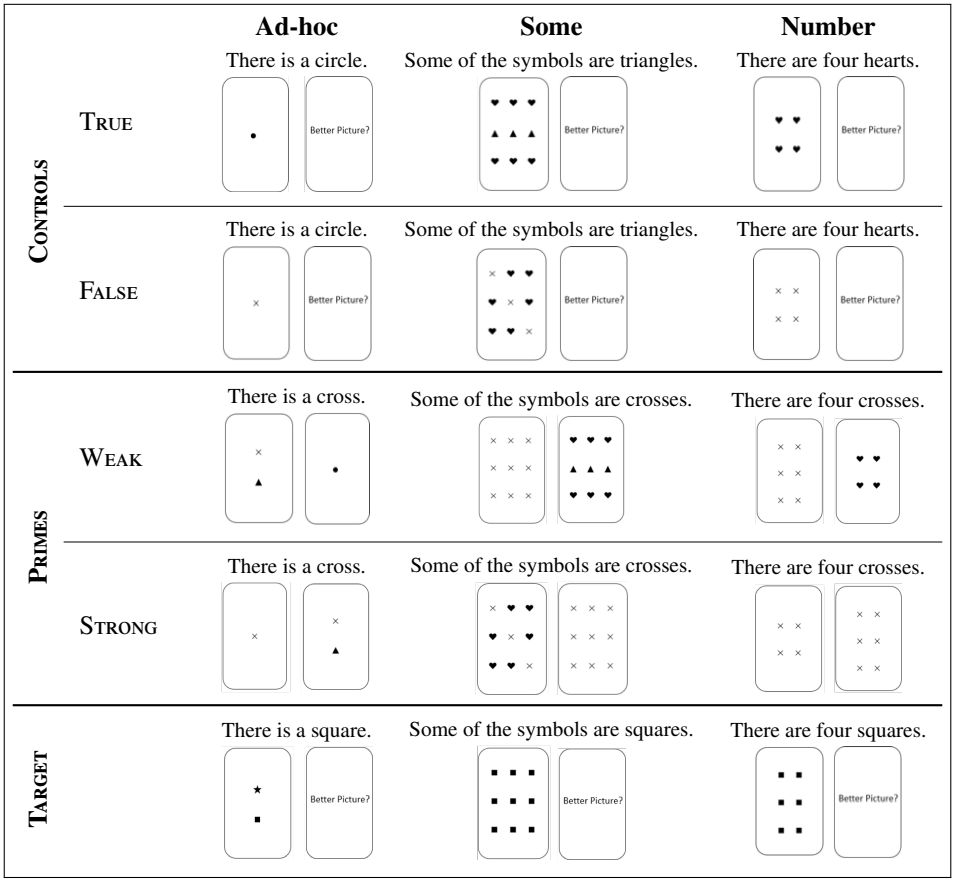


Figure 2: Example control, priming and target trials for each expression type. In the priming trials, participants choose the symbol card that best fits the sentence (the expected choice corresponds to the left card). In the control and target trials, the choice is between a symbol card and the covered card. In the target trials, if participants interpret the sentence as conveying an exhaustive meaning, they should select the covered card; otherwise, they should select the symbol card.

single symbol: on strong cards, this symbol matched the [symbol] term whereas, on false cards, it didn't. For *Some* trials, weak cards involved nine symbols of the type that matched the [symbol] term. Strong and false cards contained nine symbols, three symbols of one type and six symbols of another type: on strong cards, the minority symbol type matched the [symbol] term whereas, on false cards, none of the symbols did. Finally, for *Number* trials, weak cards contained six symbols that matched the [symbol] term. Strong and false cards contained four symbols: on strong cards, these symbols matched the [symbol] term whereas, on false cards, they didn't. As a result, for each expression type, the symbol cards in the priming trials were configured in a similar fashion in the Weak and in the Strong priming trials (i.e., all priming trials involved one card in the strong card configuration and one card in the weak card configuration).

This parallelism ensured that, in the primed target trials, participants were aware of the possibility of a strong card configuration existing behind the covered card.

Following Bott & Chemla (2016), we included filler trials related to each expression type. These trials involved alternative sentences which, relative to the weak cards, were more informative than the sentences used in the experimental trials. There were conjunctive sentences of the form *There is a [symbol] and a [other symbol]* (an alternative to *ad-hoc* sentences), sentences with the quantifier *all* (an alternative to *some*) of the form *All of the symbols are [symbol]*, and sentences with the higher numeral *six* (an alternative to *four*) of the form *There are six [symbol]*. These sentences were presented with one of three card configurations: (1) a weak card consistent with the sentence being true and the covered card, (2) a strong card inconsistent with the sentence being true and the covered card, or (3) a weak and a strong card. As in Bott & Chemla (2016), these filler trials served two purposes. First, these trials were used to help participants imagine an alternative way of describing the uncovered, weak cards involved in the target trials and to ensure that they maintained an awareness of such alternatives throughout the task. Second, these trials were used to avoid that the 'prime-prime-target' priming configuration be repeated too many times in a row, preventing the participants from recognizing the pattern and adopting response strategies.

For each experimental and filler trial, the symbol type used in the sentence was picked at random from our list of symbol types, with replacement across trials. The contents of the symbol cards accompanying each sentence were pseudo-randomly determined according to the relevant expression and the relevant condition: the matching symbol type (for strong and weak cards) always corresponded to the symbol type used in the sentence, and the non-matching symbol types were randomly chosen from our list by excluding the matching symbol type. For each trial, the position of the two cards on the screen (left or right) was chosen randomly.

3.2.3. Design

There were two blocks of trials in the experiment, a first block in which Target trials were unprimed (Baseline conditions) and a second one in which Target trials were primed (Strong and Weak priming conditions).

Block 1 consisted of True control, False control and unprimed Target trials, all of which involved an uncovered card (strong, false or weak) and the covered card. The rationale for presenting participants with unprimed Target trials and separating these trials from subsequent priming trials was to establish, for each expression type, a baseline rate of meaning enrichment which can be used to assess the direction of the within-expression priming effects previously observed in the literature. We tested all three expression types (Ad-hoc, Some and Number) in all three conditions (True control, False control and Baseline), with four iterations of each condition, giving rise to 36 experimental trials. Block 1 further included 12 filler trials (4 trials per expression type), all of which involved an uncovered card and the covered card, just like experimental trials. Half of them were presented at the very beginning of Block 1 and the other half were interspersed with the experimental trials. For half of them, the correct response corresponded to the uncovered card; for the other half, it corresponded to the covered card. Thus, Block 1 included a total of 48 individual trials.

Block 2 included the primed Target trials and were modelled after the within-expression trials from Bott & Chemla (2016). Specifically, each Target trial in Block 2 was directly preceded by either two Strong or two Weak priming trials. We tested all three expression types in both priming conditions (Strong and Weak), with four iterations of each condition, giving rise to 72 triplets, each triplet comprises two prime trials and one target trial. Block 2 further included 18 filler trials, 6 per expression type with equal numbers of the three filler types described above. Thus, the filler trials in Block 2 involved either an uncovered card and the covered card, as in the Target trials, or two uncovered cards, as in the priming trials. Half of them were presented at the very beginning of Block 2 and the rest was interspersed with the experimental triplets. Thus, Block 2 included 72 triplets of prime-prime-target and 18 individual filler trials.

3.2.4. Procedure

The experiment was run as an online survey. The survey had two parts, one for each block of experimental trials, with a self-timed break in between. Participants were given general instructions at the beginning of the survey and they were then given more specific instructions before starting each part (see Appendix Appendix A).

In the first part, participants were told that they would be presented with sentences, and that each of them would be accompanied by two pictures, one visible to them and another one covered with the text 'Better Picture?' on it. They were instructed to click on the visible picture if they considered it a match for the sentence, otherwise to click on the covered picture. Following these instructions, the experiment started with the trials from Block 1 (Baseline conditions). In the second part, participants were told that, in some cases, both pictures would now be visible to them and they were instructed to click on the picture that they considered a better match for the sentence. Following these instructions, the experiment proceeded with the trials from Block 2 (Strong and Weak priming conditions).

Each block started with some filler trials (see Design above), allowing participants to get familiar with the visual display and response procedure before they see experimental trials. All following trials (individual trials or triplets) were presented in random order in each block. On each trial, a fixation cross appeared and remained on the screen for 500 ms before the items were displayed. For each item, participants provided their response by clicking with the mouse on the picture of their choosing. Items remained on the screen until participants gave their response.

3.3. Results

3.3.1. Data treatment

Only responses to experimental trials were considered for data treatment and analyses. Responses from 1 participant were excluded from analyses because their performance to Control trials in Block 1 did not reach the threshold of 80% accuracy we had pre-established. The mean accuracy rate of the remaining participants was 97.7% (95%CI = [96.2, 98.6]) for the True control trials and 96.8% (95%CI = [95.1, 97.9]) for the False control trials. Next, following Raffray & Pickering (2010) and Bott & Chemla (2016), we removed all responses to primed Target trials that were not preceded by the two correct prime responses. In total, 138 out of 1,320 responses to primed Target trials

were removed due to incorrect prime responses (about 10% of the primed Target trials, 7% of all Target trials and 4% of the whole data set).

3.3.2. Data analyses

We analysed the data by modeling response-type likelihood using logit mixed-effects regression models (Jaeger, 2008). Analyses were conducted using the lme4 (Bates, Maechler & Bolker, 2011; Bates, Mächler, Bolker & Walker, 2014) libraries for the R statistics program (R Core Team, 2021). Analyses primarily aimed at assessing for each expression type whether responses to the Target trials in the Strong priming condition differ from those in their corresponding Baseline and Weak priming conditions. For each expression type, the model included Condition as a fixed effect (3 levels: Baseline, Strong and Weak) and the maximal random effect structure justified by the design and supported by the data, as recommended by Barr, Levy, Scheepers & Tily (2013). In the present case, the maximal converging models included random intercepts for Subject and, in some instances, random slopes for Condition grouped by Subject. In the models, Strong was first set as a reference level and compared to the other two conditions using dummy coding. The comparison between Weak and Strong permits to establish whether there is a priming effect while the comparison between Baseline and Strong permits to determine the direction of this effect, if there is one, as well as to investigate whether participants' inclination towards strengthened readings was affected in one way or another in the course of the experiment. The remaining comparisons between Baseline and Weak were conducted by changing the reference level to Weak in the models. In a similar vein as above, these comparisons were performed to investigate whether participants' inclination towards literal readings was modulated in the course of the experiment. β values, standard errors, Z-values and p-values from the lmer analyses are reported in Table 1. Wherever relevant, p-values were adjusted using the Bonferroni correction method for multiple testing. Concretely, because 9 comparisons were conducted, only p-values below 0.0055 were treated as significant.

3.3.3. Analyses

Figure 3 shows the proportion of 'Better Picture?' selection on Target trials for each expression type as a function of the experimental condition (Baseline, Strong and Weak). On our linking assumption, the proportion of 'Better Picture?' selection on Target trials indicates the extent to which participants interpreted the sentence in these trials as conveying an exhaustive meaning. For simplicity, we will refer to this measure as the rate of pragmatic responses.

Model results are shown in Table 1. All the comparisons between Strong and Weak conditions showed a priming effect in the expected direction: for each expression type, participants gave significantly less pragmatic responses in the Weak than in the Strong priming conditions (all βs < -2.2, all ps $< .001$, all adjusted ps $< .001$). These results replicate the within-expression contrasts between Strong and Weak priming conditions previously found in the literature (Bott & Chemla, 2016; Rees & Bott, 2018). By contrast, the comparisons between Baseline and priming conditions yielded a different outcome for each expression type.

For Ad-hoc, the rate of pragmatic responses in the Baseline conditions was very low ($M = 11\%$, 95%CI = $[8, 16]$), significantly lower than in the Strong conditions

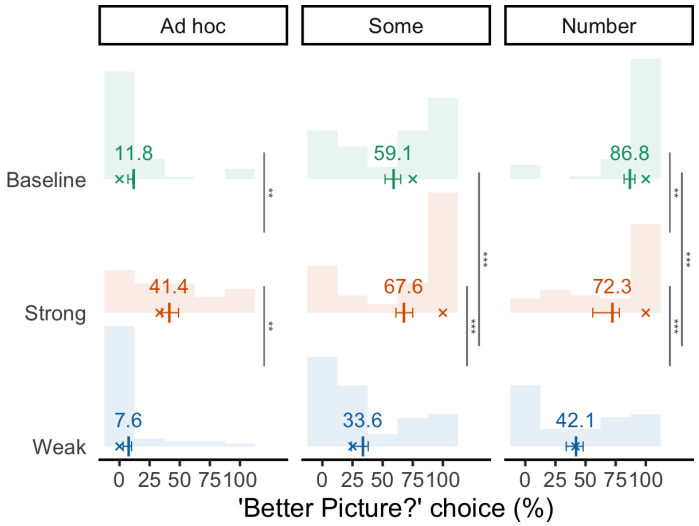


Figure 3: Proportion of 'Better Picture?' selection on target trials in Experiment 1 by expression type and condition. For each condition, the distribution of by-participant mean proportions is visualised by a histogram, the grand mean by a thick bar with its value on top and the 95% CI around it, and the median by a cross. The significance levels are based on the adjusted p-values for all comparisons tested.

		β	S.E.	Z	p-value	Adjusted
Ad-hoc						
	Strong vs. Weak	−3.271	0.422	−7.736	< .001	< .001
	Strong vs. Baseline	−2.570	0.357	−7.192	< .001	< .001
	Weak vs. Baseline	0.701	0.414	1.691	.09	0.8
Some						
	Strong vs. Weak	−2.837	0.626	−4.527	< .001	< .001
	Strong vs. Baseline	−0.921	0.705	−1.305	.19	1
	Weak vs. Baseline	1.916	0.489	3.919	< .001	< .001
Number						
	Strong vs. Weak	−2.218	0.317	−6.986	< .001	< .001
	Strong vs. Baseline	1.574	0.341	4.610	< .001	< .001
	Weak vs. Baseline	3.792	0.400	9.481	< .001	< .001

Table 1: Outputs of the Generalized linear mixed-effects models used to analyse participants' responses to target trials in Experiment 1. Adjusted p-values are provided in the last column.

(M = 42%, 95%CI = [35, 49]) and about the same as in the Weak conditions (M = 7%, 95%CI [4,11]). These results show that participants started the experiment with a clear preference for the literal reading of *Ad-hoc* sentences and that the priming effect found for these sentences is due to above-baseline rates after Strong primes. For Number, on the other hand, the rate of pragmatic responses in the Baseline conditions was very high (M = 86%, 95%CI = [81, 90]) and significantly higher than those found in the

Strong (M = 72%, 95%CI = [65, 77]) and the Weak priming conditions (M = 41%, 95% CI = [35, 48]). These results are thus the mirror image of those observed for Ad-hoc: participants initially favored the strengthened reading of *Number* sentences and the priming effect found for these sentences is due to below-baseline rates after Weak primes. The finding of below-baseline rates after Strong primes also suggests that participants' initial preference towards the strengthened reading of *Number* sentences was somewhat lessened in the course of the experiment. Finally, for Some, the rate of pragmatic responses in the Baseline conditions was intermediate (M = 59%, 95%CI = [52, 65]), with a sizeable amount of variability in responses between participants. Compared to the Baseline conditions, participants gave significantly less pragmatic answers in the Weak conditions (M = 31%, 95%CI = [25, 38]). No significant difference between Baseline and Strong conditions was found, despite a small increase of pragmatic responses in the latter (M = 68%, 95%CI = [61, 74]). Taken at face value, these results suggest that the priming effect found for Some is due to below-baseline rates after Weak primes.

However, we note that, in contrast to what we found for Ad-hoc and Number, the baseline results for Some did not exhibit any general preference for one reading type over the other. Rather, they suggest that there were two distinct profiles of participants in this case, those who favored the strengthened reading and those who favored instead the literal reading (see the top cell of the second panel in Figure 3). If priming effects are inverse preference effects, then priming effects going in opposite directions may in fact co-exist in the results for Some. That is, it is possible that, for those participants whose baseline preference is the literal reading, Strong primes had in fact a sizeable boosting effect, but this effect is concealed in the aggregate data. To explore this possibility, we conducted post-hoc analyses.

3.4. Post-hoc analyses

The baseline results for Some were more intricate than those for Ad-hoc and Number in showing a larger amount of variability in responses between participants. Specifically, the histogram in Figure 3 show that, for Some, there are two peaks in the distribution of participants' baseline rates, suggesting the presence of more than one mode. To verify this hypothesis, we first tested for unimodality of the distribution of the by-participant mean rates in the Baseline conditions by calculating the Hartigan dip-test statistic (Hartigan & Hartigan, 1985) in R diptest package (Maechler, 2013), with the alternative hypothesis that the distribution was multimodal. Results showed that the baseline rates were distributed unimodally for Ad-hoc and Number, but not for Some (D = 0.112, adjusted $p < .001$). Taking the assumption of two modes being present in the baseline data for Some, we next estimated the location of these modes and their density value. The mode with the highest estimated density value peaked at 99% and the second one at 1%, with the antimode identified at 47%. These values indicate that some participants consistently understood the *Some* sentences in the Baseline conditions under the pragmatic reading, while others consistently understood them under

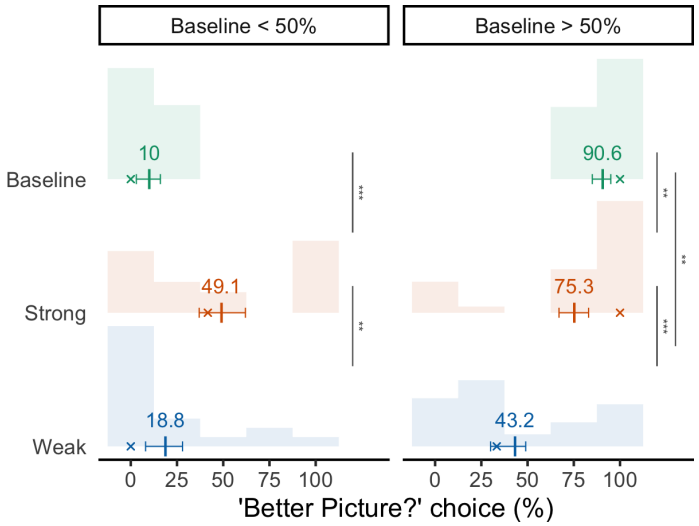


Figure 4: Proportion of 'Better Picture?' selection on target trials for SOME in Experiment 1, sorted according to whether the rate in the BASELINE condition is above or below 50%. For each condition, the distribution of by-participant mean proportions is visualised by a histogram, the grand mean by a thick bar with its value on top and the 95% CI around it, and the median by a cross. The significance levels are based on the adjusted *p*-values for all comparisons tested.

the literal reading.[6]

In order to see how these two groups of participants were affected by the Weak and Strong primes, we sorted the results of the *Some* target trials according to two responder profiles: participants were classified as Literal-Some responders if their baseline rate was below 50% and as Pragmatic-Some responders if their baseline rate was above 50%. Following this partition, there were 32 Pragmatic-Some responders and 20 Literal-Some responders, representing 58% and 36% of the subjects in our sample, respectively. Three participants had a baseline rate of exactly 50%. Their results were not included in the following analyses. Figure 4 shows the proportion of pragmatic responses on *Some* target trials by responder group and experimental condition. Responses from both groups were analyzed using the data analysis pipelines from the main analyses, i.e., by conducting pairwise comparisons of all three conditions. The maximal converging models included a random intercept for Subject. Model results are shown in Table 2. The *p*-values reported here are adjusted based on the total number of comparisons conducted in the main and the post-hoc analyses, which amounted to lowering the significance level to 0.003.

[6]This finding aligns with results from previous studies reporting substantial variation in responses to underinformative scalar sentences and a bimodal distribution between pragmatic and logical responders with adults (Teresa Guasti, Chierchia, Crain, Foppolo, Gualmini & Meroni, 2005; Noveck & Posada, 2003; Hunt, Politzer-Ahles, Gibson, Minai & Fiorentino, 2013) as well as with children (Noveck, 2001; Teresa Guasti et al., 2005; Foppolo, Guasti & Chierchia, 2012; Horowitz, Schneider & Frank, 2018; Foppolo, Mazzaggio, Panzeri & Surian, 2021).

		β	S.E.	Z	p-value	Adjusted
Literal-Some						
	Strong vs. Weak	−1.809	0.504	−3.583	< .001	< .01
	Strong vs. Baseline	−2.723	0.553	−4.920	< .001	< .001
	Weak vs. Baseline	−0.913	0.550	−1.661	.09	1
Pragmatic-Some						
	Strong vs. Weak	−1.984	0.366	−5.418	< .001	< .001
	Strong vs. Baseline	1.467	0.428	3.427	< .001	< .01
	Weak vs. Baseline	3.451	0.459	7.505	< .001	< .001

Table 2: Outputs of the Generalized linear mixed-effects models used in the post-hoc analysis of the *Some* target trials in Experiment 1. Adjusted p-values are provided in the last column.

For both groups, the comparison between Strong and Weak showed a priming effect so that both Literal-Some and Pragmatic-Some responders gave significantly less pragmatic responses in the Weak than in the Strong conditions (all βs < -1.8, all ps $< .001$, all adjusted ps $< .001$). Crucially, however, the comparisons between Baseline and priming conditions showed that these priming effects were driven by a different prime type in each case. For the Pragmatic-Some responders, the rate of pragmatic responses in the Baseline conditions ($M = 90\%$, 95%CI = $[84, 94]$) was significantly higher than those found in the Strong ($M = 75\%$, 95%CI = $[67, 82]$) and the Weak conditions ($M = 39\%$, 95%CI = $[31, 49]$). In other words, for these subjects, the presentation of Weak primes had a sizeable inhibition effect on pragmatic responses, consistent with what we observed in the main analyses. However, for the Literal-Some responders, the rate of pragmatic responses in the Baseline conditions ($M = 10\%$, 95%CI = $[5, 18]$) was about the same as in the Weak conditions ($M = 18\%$, 95%CI = $[10, 29]$), but significantly lower than in the Strong conditions ($M = 50\%$, 95%CI = $[37, 62]$). These results indicate that the priming effect found for this group of subjects was driven by the presentation of Strong primes, which had a boosting effect on pragmatic responses. Taken together, these findings establish the co-existence of two types of priming effects in the results for Some, one driven by the Strong primes and the other by the Weak primes. More generally, these results suggest that the opposite priming effects we found for Literal-Some and Pragmatic-Some responders are entirely parallel to those we found for Ad-hoc and Number in the main analyses.

3.5. Discussion

Our results replicate the within-scale priming effects from Bott & Chemla (2016) in full in showing that, for all three scalar expressions we tested, participants systematically provided more responses based on SIs after Strong priming trials than after Weak priming trials. In addition, our original finding is that these seemingly homogeneous priming effects are in fact driven by distinct prime types depending on speakers' interpretive preferences prior to being exposed to primed trials. Specifically, the novel Baseline conditions we introduced inform us that, for Ad-hoc sentences, Strong priming trials boosted pragmatic responses, the less frequent response outcome prior to the priming phase. For Number sentences, however, no such boosting effects were observed. Rather, for these items, we found that Weak priming trials inhibited the more

frequent pragmatic interpretation in favor of the less frequent literal one, by giving rise to below-baseline rates of pragmatic responses. Finally, for SOME sentences, both types of priming effects were found to co-exist in our data, with a principled distribution across participants: for the literal-some responders, Strong primes had a boosting effect on pragmatic responses, parallel to the one observed for Ad-hoc; by contrast, for the pragmatic-some responders, Weak primes had an inhibition effect, parallel to the one we found for Number.

These findings show that, contrary to what Bott & Chemla (2016) suggest, what is primed in this experimental set-up is not necessarily the generation of SIs or, similarly, the enriched interpretation: in the case of Number, as well as Some for the pragmatic responders, what is primed is instead the suspension of the relevant SIs and, consequently, the literal, un-enriched interpretation. Our findings also show that the direction of within-scale priming effects is not random. Rather, the direction of these effects appears to be predictable from speakers' prior preferences: they are all driven by the prime type favoring the less preferred interpretation prior to priming, as established by the baseline results. In the case of Ad-hoc, as well as Some for literal responders, the less preferred interpretation corresponded to the one with SIs, which was thus promoted by the Strong primes; for Number, as well as Some for pragmatic responders, this corresponded to the literal interpretation, the one without SIs, which was thus promoted by the Weak primes. Taken together, these findings support the view that within-expression priming effects are *inverse preference effects*.

Finally, the present results suggest that certain priming trials had wider-ranging effects on decisions, i.e., beyond the immediate triplet of interest. Specifically, in the case of Number, as well as Some for pragmatic responders, we found that the rates of pragmatic responses after Strong primes were in fact lower than the baseline rates. Given that the enriched interpretation was initially favored in such cases, these contrasts suggest that the Weak primes favoring the less frequent literal interpretation also affected participants' responses to the Strong priming trials. That is, not only did the Weak primes promote the literal interpretation in the Weak priming trials, their presence may also have caused 'spill-over' effects in promoting this interpretation in the Strong priming trials as well, hence the below-baseline rates observed for these trials. We immediately note, however, that the results of Experiment 1 are not conclusive regarding the presence of spillover effects. In particular, no such effects were observed in the case of Ad-hoc, or in the case of Some for the literal responders: in these cases, the Strong primes driving the priming effects did not seem to have affected participants' responses to the Weak priming trials.

In principle, the absence of Strong spillover effects in our data could indicate that these effects are weaker, and thus harder to detect than their Weak counterparts. It could be so for instance if Strong primes have a lesser priming potential than Weak primes, e.g., because it is harder to prime the computation of an SI than to prime its suspension. This option would align well with the observation that, in our data, Strong priming generally induced smaller changes in baseline preferences compared to Weak priming. In the absence of evidence for Strong spillover effects, however, the contrasts of interest leave room for other explanations. For instance, these contrasts could indicate that, as the experiment progressed, participants provided less pragmatic responses in the Strong conditions of certain items (e.g., Number) in order to compensate for the increase of

pragmatic responses in the Strong conditions of others (e.g., Ad-hoc), possibly in an attempt to mitigate the processing effort associated with SI generation. As far as we can see, such a trade-off strategy would similarly explain the observed contrasts.

Now, recall that, following Bott & Chemla (2016), all filler items in Experiment 1 involved canonical scalar alternatives to the test sentences, part of which was presented at the start of each block of trials. As we explained, one of the motivations for including such items was to help participants imagine an alternative way of describing the weak cards used in the target trials. In light of the present discussion, however, it bears pointing out that these items may also have had a broader impact on participants' reasoning: by increasing awareness of alternatives throughout the task, these items may have decreased participants' sensitivity to the Strong primes while increasing their sensitivity to the Weak primes. Thus, the presence of these fillers may have magnified certain contrasts between baseline and priming conditions while preventing others from being detected. To address this concern, we set out to refine the design of Experiment 1 by using more neutral fillers while reducing their number to the minimum required by our experimental purposes.

4. Experiment 2: Testing the Effect of Alternative Fillers

Results from Experiment 1 indicate that within-expression priming effects are inverse preference effects: in all the cases we surveyed, we found that the most effective prime type was the one promoting the less favored interpretation of a given expression, as established from participants' prior preferences. As we discussed, however, these results leave open the question of whether or not effective prime types induce wider-ranging, spillover effects in promoting less canonical interpretations across-the-board, i.e., across priming conditions. The purpose of Experiment 2 was to investigate this question further by retesting the same test items as in Experiment 1 but, this time, without presenting participants with more informative alternative sentences at any point of the experiment.

4.1. Data availability

Stimuli, data and analysis code associated with Experiment 2 are available open access on the OSF at https://osf.io/263xf/.

4.2. Methods

4.2.1. Participants

60 novel participants (38 female; average age 34.4 yrs) were recruited online through Prolific using the same pre-screening criteria as in Experiment 1. Of these, 1 was excluded prior to data treatment because they did not declare English as their native language in our demographic survey. Participants were paid £1.30 and average completion time was about 8 minutes. The consent and data collection procedures were the same as in Experiment 1.

4.2.2. Materials and Design

The materials were the same as in Experiment 1 except for the filler trials. In contrast to Experiment 1, these trials involved the same sentence types as those used in the test trials (see the frames in (3) above). Sentences in these trials were paired with one of three card configurations: (1) a weak and a false card, as in the Weak priming trials, (2) a strong and a weak card, as in the Strong priming trials, or (3) a strong card and the covered card, as in the Target trials. For each scalar expression, there were 2 instances of the first filler type, 2 instances of the second and 4 instances of the third. Thus, there were 8 fillers per expression, half of them were virtually identical to priming trials, the other half to target trials. Filler trials were used in Block 2, and only in Block 2, with the sole purpose of preventing participants from identifying the 'prime-prime-target' pattern of the experimental triplets. These items were generated through the same randomization procedure as the priming and test trials.

The rest of the design was identical to that of Experiment 1 so that Figure 2 also stands as an illustrative summary of the control, priming and test trials used in Experiment 2. Block 1 consisted only of control and unprimed Target trials (no fillers). As in Experiment 1, all three expression types were tested in the True, False and Target conditions, with four iterations of each condition, giving rise to 36 individual trials. Priming and primed trials were presented in Block 2, together with the novel fillers we just described. As in Experiment 1, all three expressions were tested in both priming conditions, with four iterations of each condition. Thus, Block 2 included 72 triplets of test trials and 24 individual filler trials (8 per expression), which were interspersed with the test triplets.

4.2.3. Procedure

The procedure and instructions were the same as in Experiment 1 (see Appendix A for the instructions).

4.3. Results

4.3.1. Data treatment

All participants' performance to Control trials in Block 1 reached the pre-established threshold of 80% accuracy. Data from all the participants were thus included in our analyses. Participants' mean accuracy rate was 98.4% (95%CI = $[97.2, 99.1]$) for the True control trials and 95.9% (95%CI = $[94.1, 97.1]$) for the False control trials. Following the same procedure as in Experiment 1, we removed all responses to primed Target trials that were not preceded by the two correct prime responses. In total, 154 out of 1,416 responses to primed Target trials were removed due to incorrect prime responses (about 10% of the primed Target trials, 7% of all Target trials and 4% of the whole data set).

4.3.2. Data analyses

Responses to test trials were analysed using the data analysis pipelines from Experiment 1. Based on the findings from Experiment 1, we made a preliminary test for unimodality of the distribution of the by-participant mean rates in the Baseline conditions. Results were similar to those from Experiment 1 in showing that the baseline rates were

distributed unimodally for Ad-hoc and Number, but not for Some ($D = 0.169$, adjusted $p < .001$).[7] Thus, for the present analyses, results to *Some* target trials were directly sorted according to the two profiles of responders we had already identified. In total, there were 30 Pragmatic-Some responders and 26 Literal-Some responders, representing about 50% and 44% of the subjects in our sample, respectively.[8] Accordingly, pairwise comparisons between Baseline, Strong and Weak conditions were conducted for Ad-hoc, Number, Literal-Some responders and Pragmatic-Some responders. The models included Condition as a fixed effect and a random intercept for Subject. As before, p-values were adjusted using the Bonferroni correction method for multiple testing.

4.3.3. Analyses

Figure 5 shows the proportion of 'Better Picture?' selection on Target trials (i.e., the rate of pragmatic responses) by condition for Ad-hoc and Number (top row), and by condition and responder group for Some (bottom row).

Model results are shown in Table 3. In line with the results from Experiment 1, all the comparisons between Strong and Weak conditions showed a priming effect in the expected direction (all βs < -1.08, all ps $< .01$). After correction for multiple comparisons, the effect was significant for Ad-hoc, Literal-Some and Number (adjusted ps $< .01$) and marginally significant for Pragmatic-Some (adjusted $p = .07$). As before, the comparisons between Baseline and priming conditions were used to identify the prime types driving these effects and probe for potential wider-ranging effects. For both types of effects, Ad-hoc and Literal-Some were found to pattern alike, and differently from Number and Pragmatic-Some.

For Ad-hoc, the rate of pragmatic responses in the Baseline conditions ($M = 8\%$, 95%CI = $[5, 12]$) was significantly lower than in the Strong conditions ($M = 33\%$, 95%CI = $[27, 39]$, adjusted $p < .001$), but also than in the Weak conditions ($M = 17\%$, 95%CI = $[12, 22]$, adjusted $p < .01$). Thus, these results replicate those from Experiment 1 in establishing that the priming effect for Ad-hoc is essentially driven by the Strong primes, which substantially increased pragmatic responses compared to the baselines. In addition, they show that, in contrast to what we found in Experiment 1, the Weak priming conditions also yielded above-baseline rates, suggesting that the pragmatic interpretation of *Ad-hoc* sentences was promoted by the Strong primes across priming conditions. Similar effects were found in the results for Some with the Literal-Some responders. For these participants, the baseline rate pragmatic responses ($M = 5\%$, 95%CI = $[2, 12]$) was significantly lower than in the Strong conditions ($M = 45\%$, 95%CI = $[35, 56]$, adjusted $p < .001$) as well as marginally lower than in the Weak conditions ($M = 18\%$, 95%CI = $[11, 29]$, adjusted $p = .06$). Thus, in a way similar to what we found for Ad-hoc, the Strong primes driving the priming effect for Some among Literal-Some responders seem to have promoted the pragmatic

[7] As in Experiment 1, there appeared to be two modes in the baseline data for Some (see also Figure 5). The mode with the highest estimated density value peaked above 99% and the second one below 1%, with the antimode identified at 43%.

[8] Three participants had a baseline rate of exactly 50%, and so their results were not included in the analyses of the Some trials.

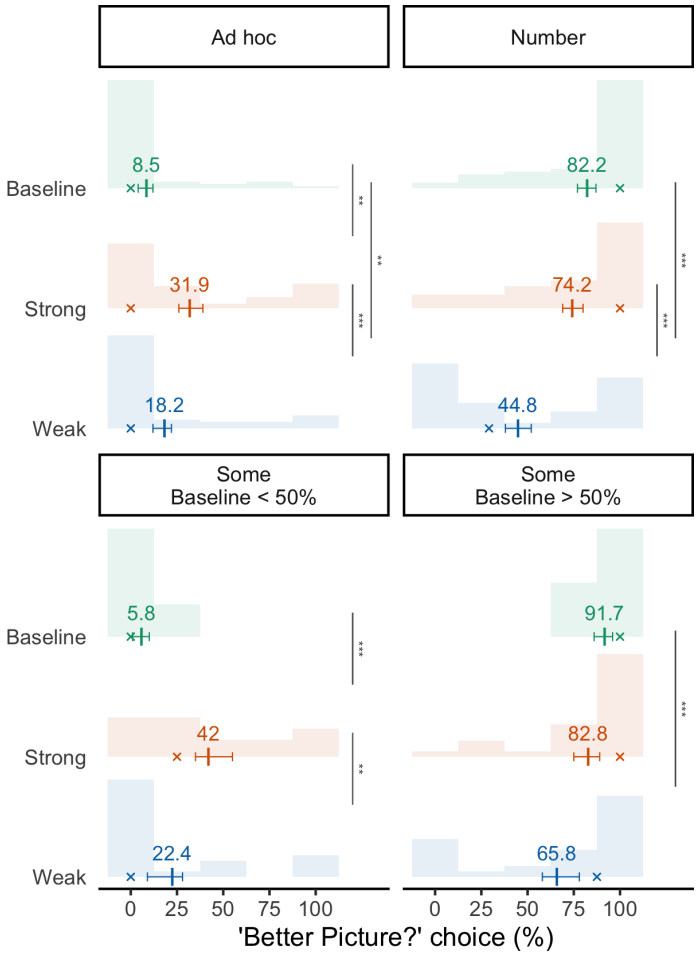


Figure 5: Proportion of 'Better Picture?' selection on target trials in Experiment 2 by condition for Ad-hoc and Number (top row) and by condition and responder group for Some (bottom row). For each condition, the distribution of by-participant mean proportions is visualised by a histogram, the grand mean by a thick bar with its value on top and the 95%CI around it, and the median by a cross. The significance levels are based on the adjusted p-values for all comparisons tested.

interpretation not just in the Strong priming, but also in the Weak priming conditions.

By contrast, the priming effects for Number and Pragmatic-Some responders were driven by the Weak primes, exactly as in Experiment 1. For Number, the rate of pragmatic responses in the Baseline conditions (M = 82%, 95%CI = [76, 86]) was significantly higher than in the Weak conditions (M = 45%, 95%CI = [38, 52], adjusted $p < .001$) and slightly higher than in the Strong conditions (M = 74%, 95%CI = [69, 80], adjusted $p = .17$), although the latter contrast did not remain significant after correction. These results replicate the effects observed for Number in Experiment 1 in showing that exposure to the Weak primes decreased participants' pragmatic responses

		β	S.E.	Z	p-value	Adjusted
Ad-hoc						
	Strong vs. Weak	−1.524	0.343	−4.444	< .001	< .001
	Strong vs. Baseline	−3.044	0.429	−7.095	< .001	< .001
	Weak vs. Baseline	−1.519	0.408	−3.719	< .001	< .01
Literal-Some						
	Strong vs. Weak	−1.593	0.452	−3.519	< .001	< .01
	Strong vs. Baseline	−3.217	0.564	−5.702	< .001	< .001
	Weak vs. Baseline	−1.623	0.576	−2.815	< .01	.06
Pragmatic-Some						
	Strong vs. Weak	−1.083	0.399	−2.715	< .01	.07
	Strong vs. Baseline	1.012	0.451	2.244	< .05	.29
	Weak vs. Baseline	2.096	0.470	4.453	< .001	< .001
Number						
	Strong vs. Weak	−2.253	0.304	−7.407	< .001	< .001
	Strong vs. Baseline	0.694	0.284	2.440	< .05	.17
	Weak vs. Baseline	2.947	0.329	8.955	< .001	< .001

Table 3: Outputs of the Generalized linear mixed-effects models used to analyse participants' responses to target trials in Experiment 2. Adjusted p-values are provided in the last column.

not only in the Weak priming conditions, but also in the Strong priming conditions, albeit to a lesser degree. Similarly, for Pragmatic-Some, the rate of pragmatic responses in the Baseline conditions (M = 91%, 95%CI = [85, 95]) was significantly higher than in the Weak conditions (M = 68%, 95%CI = [57, 77], adjusted $p < .001$) and slightly higher than in the Strong conditions (M = 82%, 95%CI = [74, 88], adjusted $p = .29$), although the latter contrast did not remain significant after correction. These results suggest that, in the case of Some, exposure to Weak primes only mildly affected the responses of Pragmatic-Some responders in the Strong priming conditions.

4.4. Discussion

The goal of Experiment 2 was to assess the generality of the various effects observed in Experiment 1 by retesting the same test items in a more controlled experimental environment. Specifically, Experiment 2 aimed to control for potential effects associated with the use of more informative sentences in the fillers of Experiment 1. For these purposes, we modified the design of Experiment 1 by reducing the number of filler items and, more importantly, by altering their contents so that participants were no longer presented with canonical scalar alternatives to the test sentences during the experiment. In bringing these modifications, we hoped to obtain more fine-grained comparisons between baseline and priming conditions which we could use in turn to establish whether, in addition to driving priming effects, effective prime types also induce spillover effects, as suggested by the results from Experiment 1.

First, our results replicate the main findings from Experiment 1 in showing that within-expression priming effects are driven by the prime type promoting the less preferred interpretation prior to priming, consistent with the view that these effects are

inverse preference effects. Second, these results reveal a novel type of spillover effects, absent from Experiment 1. In the case of Ad-hoc, as well as Some for the literal responders, the rates of pragmatic responses were higher in the Weak priming conditions than in the baselines. These results suggest that, in both cases, exposure to the Strong primes in Block 2 promoted the less frequent pragmatic interpretation beyond the Strong priming trials. Finally, we note that, as in Experiment 1, the results for Number showed spillover effects in the opposite direction – that is, driven by the Weak primes – and that the results for Some showed a similar trend among the pragmatic responders, although the relevant contrasts in both these cases were less pronounced than in Experiment 1. We take these findings to provide evidence for the fact that effective prime types generally have wider-ranging effects in promoting the initially less preferred interpretation across other trials and conditions.

More generally, the discrepancies observed between Experiment 1 and Experiment 2 regarding the distribution and strength of the spillover effects show that, in priming experiments like ours, participants' responses to test items can easily be affected by related items present elsewhere in the experiment. In the present case, the relevant discrepancies can directly be attributed to the presence vs. absence of more informative sentences in the fillers of Experiment 1 vs. Experiment 2. In our view, it is possible that the presence of such items contributed their own priming effects in Experiment 1: by maintaining participants' awareness of alternatives throughout the task, these items may have facilitated scalar reasoning and, as a result, reduced participants' sensitivity to the Strong primes or, similarly, increased their sensitivity to the Weak primes, hence the absence of detectable Strong-driven spillover effects in Experiment 1.

The finding that responses to a priming condition may be affected by the presentation of other trials in the experiment is also important in interpreting previous results from the literature. Thus, for instance, in the case of Some and Number, Waldon & Degen (2020) observed an increase in the rate of pragmatic responses after Strong primes, compared to their baselines. We note however that the baseline trials in their experiment were in the same block of trials as the Weak and Strong priming trials. In light of our findings, the contrasts Waldon & Degen observed could thus be due to a spillover effect of the Weak priming trials on the baseline conditions. We believe that these considerations may invite future work on implicature priming to opt for the kind block design we developed in our studies: by testing baseline conditions in a separate block, prior to all priming trials, this design allows one to avoid any possible spillover effects of priming trials on the baselines, increasing the reliability of the comparisons between baseline and priming conditions.

Before summarising our findings, we propose to further buttress our main conclusions by addressing a potential concern about the design of our experiments, which also applies to previous studies on implicature priming. In a nutshell, the gist of the worry is that what we described as priming effects on interpretation could in fact be driven by a linguistic-independent factor: participants possibly chose more often the overt card after Weak primes simply because they had just chosen visually similar cards in the preceding priming trials; similarly, they may have chosen more often the covert card after Strong primes because the overt card didn't visually match in these cases the cards they had chosen in the preceding priming trials. As it is easy to see,

such effects of visual similarity, if present, could then account, entirely or partly, for the contrasts observed between Strong and Weak priming conditions. In the third and last experiment we report on, we demonstrate that if the linguistic stimuli are removed from the priming trials, then the target contrasts completely disappear, suggesting that visual similarity plays no role in within-expression priming effects.

5. Experiment 3: Testing the Effect of Visual Similarity

The goal of this experiment was to test whether the visual similarity (or dissimilarity) between the target cards in the priming trials and the uncovered card in the target trials could explain (part of) the effects found in Experiments 1–2 (for similar concerns, see Bott & Chemla 2016 and Rees & Bott 2018). The materials and design were the same as in previous experiments except for the contents of the priming trials: instead of involving a sentence and a choice between two symbol cards, priming trials involved no sentence and only one symbol card, which participants were forced to select. In Weak primes, the symbol card was configured like the uncovered card involved in target trials (i.e., weak card configuration). In Strong primes, the symbol card was instead configured like a strong card and thus visually different from the uncovered card involved in target trials. If participants' responses to primed target trials are driven by the visual correspondence between the priming cards and the test card, then similar contrasts between Strong and Weak priming conditions should be observed in this novel experiment.

5.1. Data availability

Stimuli, data and analysis code associated with Experiment 3 are available open access on the OSF at https://osf.io/263xf/.

5.2. Methods

5.2.1. Participants

50 novel participants (29 female, average age 35.1 years) were recruited online through Prolific using the same pre-screening criteria as in Experiments 1-2. Participants were paid £1.15, and average completion time was about 7 minutes. The consent and data collection procedures were the same as in Experiments 1-2.

5.2.2. Materials

Experimental triplets were created in a manner analogous to the experimental triplets in Experiments 1–2 with one critical difference: the priming trials involved only one symbol card and no sentence at all. Strong priming trials consisted of one symbol card in the strong card configuration: a single symbol for Ad-hoc, 3 symbols of one type and 6 symbols of another type for Some, and 4 symbols of the same type for Number. Weak priming trials consisted of one symbol card in the weak card configuration: 2 different symbols for Ad-hoc, 9 symbols of the same type for Some, and 6 symbols of the same type for Number. Symbol types in these trials were randomly chosen from our list of symbols. All other trial types (control and target) were constructed in the exact same way as their corresponding type in Experiments 1–2. Example control, priming

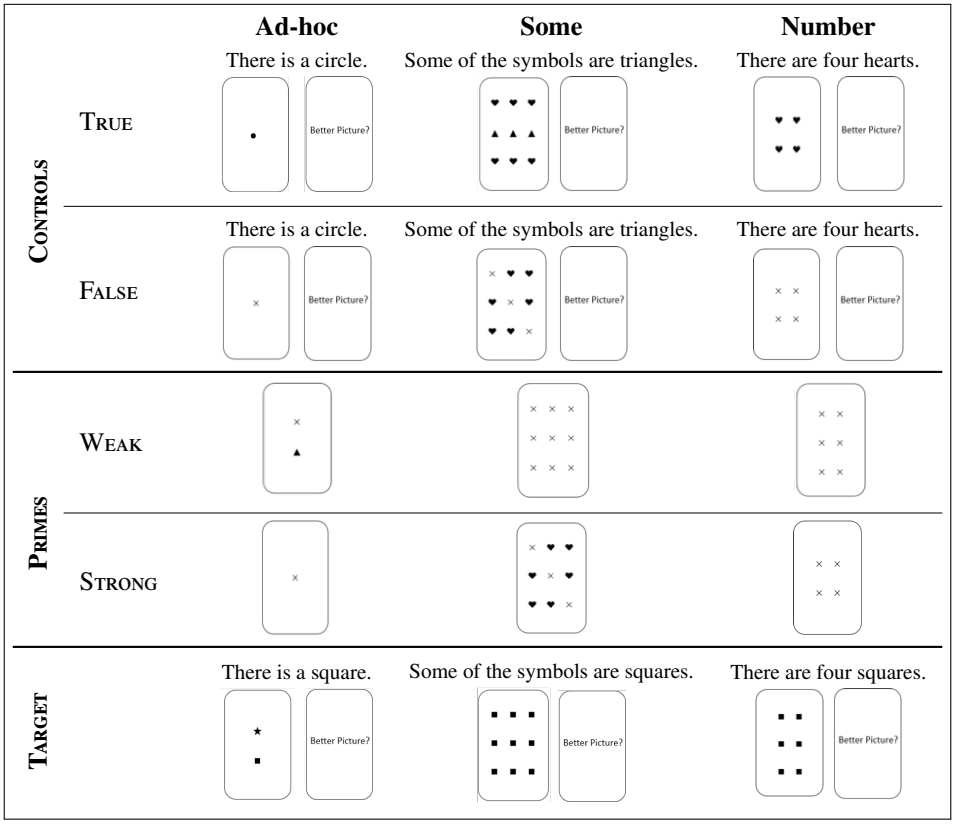


Figure 6: Example control, priming and target trials for each expression type. In the priming trials, participants selected the symbol card presented to them. In the control and target trials, they chose the card that best fits the sentence.

and target trials used in Experiment 3 are given in Figure 6. Filler trials were created in a manner analogous to the fillers in Experiment 1, again with one exception: filler trials involving a weak and a strong card were removed and replaced with filler trials involving only one symbol card and no sentence at all, just like the priming trials. The symbol types used in these trials and in the priming trials were picked at random from our list of symbol types. The rest of the design was identical to that of Experiment 1 in all regards.

5.2.3. Procedure

The procedure was the same as in Experiments 1–2 with some minor differences in the instructions given to participants (see Appendix B). Specifically, prior to completing the second block of trials, participants were told that, in some cases, they would see pages with a single picture and no sentence. They were asked in such cases to look at the picture and to click on it. Some of the filler trials placed at the start of Block 2 were instances of such cases.

5.3. Results

5.3.1. Data treatment

Responses from 1 participant were excluded because their performance to Control trials did not reach the pre-established threshold of 80% accuracy. The mean accuracy rate of the remaining participants was 97.7% (95%CI = [96.2, 98.7]) for the True control trials and 97.4% (95%CI = [95.8, 98.4]) for the False control trials. Note that, in contrast to Experiments 1–2, participants could not incorrectly respond to the priming trials. Therefore, all responses to Target trials were included in our analyses.

5.3.2. Data analyses

Data were analysed using the data analysis pipelines from Experiments 1–2. The tests of unimodality on the by-participant mean rates in the Baseline conditions yielded the same results as in Experiments 1–2: in contrast to Ad-hoc and Number, the baseline rates for Some were not distributed unimodally ($D = 0.173$, $p < .001$), with 27 participants exhibiting a strong preference for the pragmatic interpretation and 20 for the literal interpretation.[9] Thus, here again, the results to *Some* target trials were sorted according to the two groups of responders previously identified, Literal-Some and Pragmatic-Some responders. The main analyses were identical to those conducted and reported in Experiment 2.

5.3.3. Analyses

Figure 7 shows the proportion of 'Better Picture?' selection on Target trials (i.e., the rate of pragmatic responses) by condition for Ad-hoc and Number (top row), and by condition and responder group for Some (bottom row). Model results are shown in Table 4. The comparisons between Strong and Weak priming conditions didn't yield any significant contrast for any expression type or responder group (all $|\beta|$s < 0.7, all adjusted ps $= 1$). In other words, the results didn't show any of the within-expression priming effects we observed in both Experiment 1 and Experiment 2. The rates of pragmatic responses in the Baseline conditions were very similar to those found in Experiments 1–2 (around 15% for Ad-hoc, 55% for Some and 80% for Number) and no different from those in the priming conditions, with one exception: for Some, the rates of pragmatic responses among Literal-Some responders were significantly lower in the baselines ($M = 3\%$, 95%CI = $[1, 10]$) than in the Weak ($M = 16\%$, 95%CI = $[9, 25]$, adjusted $p < .01$) and the Strong priming conditions ($M = 17\%$, 95%CI = $[10, 27]$, adjusted $p < .01$). Given the absence of priming effects in this experiment, we take these results to suggest that repeated exposure to the *Some* test trials helped the Literal-Some responders access their pragmatic interpretation in the course of the experiment.

[9] The mode with the highest estimated density value peaked above 99% and the second below 1%, with the antimode identified at 40%. Two participants had a baseline rate of exactly 50%; their results were not included in the analyses of the Some trials.

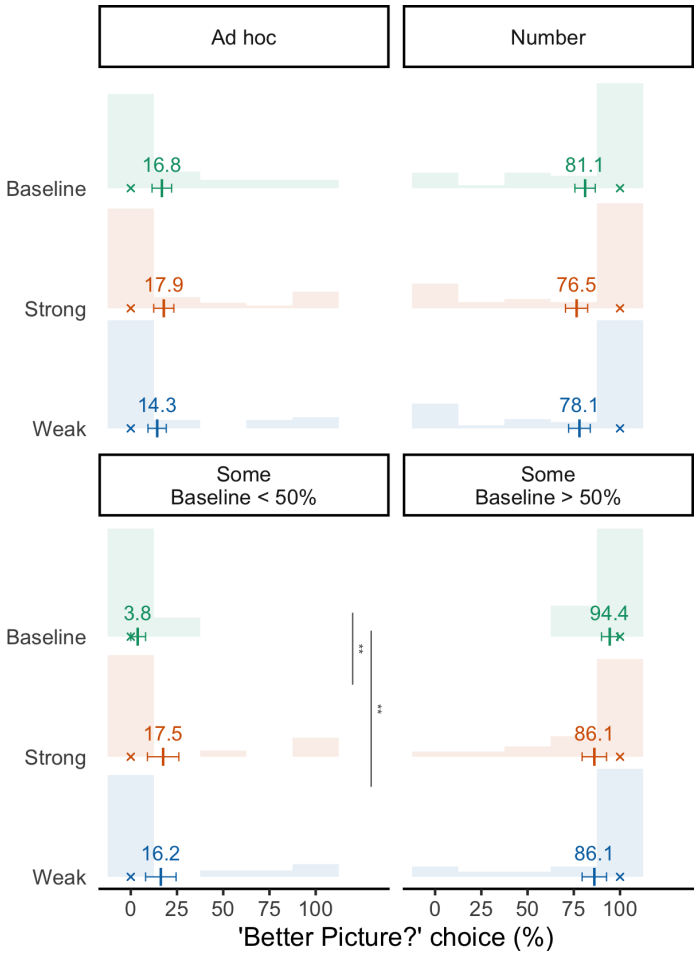


Figure 7: Proportion of 'Better Picture?' selection on target trials in Experiment 3 by condition for Ad-hoc and Number (top row) and by condition and responder group for Some (bottom row). For each condition, the distribution of by-participant mean proportions is visualised by a histogram, the grand mean by a thick bar with its value on top and the 95% CI around it, and the median by a cross. The significance levels are based on the adjusted *p*-values for all comparisons tested.

5.4. Discussion

This experiment was designed as a follow-up to Experiments 1–2 to test whether the visual similarity (or dissimilarity) between the target cards in the priming trials and the uncovered card in the target trials could explain (part of) the priming effects found in Experiments 1–2. The present results disconfirm any explanation of the priming patterns established in Experiments 1–2 solely based on the properties of our visual stimuli, thereby reinforcing our previous conclusions.

		β	S.E.	Z	p-value	Adjusted
Ad-hoc						
	Strong vs. Weak	−0.694	0.452	−1.533	.12	1
	Strong vs. Baseline	−0.188	0.434	−0.433	.66	1
	Weak vs. Baseline	0.505	0.453	1.115	.26	1
Literal-Some						
	Strong vs. Weak	−0.454	0.962	−0.473	.63	1
	Strong vs. Baseline	−3.912	1.087	−3.598	< .001	< .01
	Weak vs. Baseline	−3.457	1.021	−3.384	< .001	< .01
Pragmatic-Some						
	Strong vs. Weak	0	.506	0	1	1
	Strong vs. Baseline	1.532	0.630	2.428	< .05	.18
	Weak vs. Baseline	1.532	0.630	2.428	< .05	.18
Number						
	Strong vs. Weak	0.316	0.460	0.687	.49	1
	Strong vs. Baseline	0.949	0.470	2.017	< .05	.52
	Weak vs. Baseline	0.633	0.464	1.364	.17	1

Table 4: Outputs of the Generalized linear mixed-effects models used to analyse participants' responses to target trials in Experiment 3. Adjusted p-values are provided in the last column.

6. Inverse preference and spillover as results of context adaptation

To summarise the main findings from the three experiments, the results from the baseline blocks of Experiments 1–2 show that SOME, AD HOC and NUMBER have different baseline rates of pragmatic responses: Across participants, the weak reading of AD HOC is more likely to be accepted, while the weak reading of NUMBER is more likely to be rejected. For SOME, there is inter-speaker variation such that for about half the participants of our experiments, the baseline rate of pragmatic responses was as low as for AD HOC and for the other half, it was comparable to NUMBER.

With these baselines, we clearly see that the priming effects observed in Block 2 of Experiments 1–2 are *inverse preference effects* in the sense that priming trials that force the less preferred reading have sizable priming effects, while priming trials that force the more preferred reading hardly had priming effects.

Our experimental results also suggest that the priming effects of non-default interpretations are so strong and persistent that they had *spillover effects* on subsequent target items in other experimental conditions. Furthermore, comparing Experiments 1–2, we see that even filler items could have had such spillover effects in Experiment 1. This casts reasonable doubt on the neutrality of the baseline condition used by Waldon & Degen (2020), which was mixed with priming and target trials (and similarly for the proxy baselines of Bott & Chemla 2016).

Finally, the null results of Experiment 3, where priming trials had no linguistic material, suggest that the visual stimuli in our experiments had no priming effects, further reinforcing our conclusion that the priming effects in this experimental paradigm are inverse preference effects for the strong vs. weak readings.

6.1. Challenges for activation-based accounts

Recall that previous studies on implicature priming compared weak and strong primes and concluded that certain aspects of the mechanism of SI computation common to different kinds of SIs can be primed (Bott & Chemla, 2016; Rees & Bott, 2018; Waldon & Degen, 2020; Meyer & Feiman, 2021). In particular, Bott & Chemla (2016) discuss several analytical possibilities in detail (see also Rees & Bott 2018; Waldon & Degen 2020), and put forward a view that (at least) two types of priming are involved in implicature priming. One has to do with the use of the common mechanism of SI computation, and the other one has to do with making a particular alternative active and salient. We claim that neither of them satisfactorily explains the full inverse preference pattern we observed in our experimental results.

Firstly, Bott & Chemla (2016) observed cross-scale priming, which is priming between different scalar expressions. Specifically, they observed that the SI of a scalar item *X* was observed more often after strong primes involving a different scalar item *Y*, than after weak primes involving *Y*. In order to explain this observation, they claim that the computational mechanism common to different SIs is activated during the strong priming trials, but not during weak priming trials, and since the activation level of the mechanism stays high after strong priming trials, they are more likely to be used during the subsequent target trials following them.[10]

This type of priming on its own cannot capture the full inverse preference effects revealed in our results. In particular, it has very little to say about cases where the baseline rate of pragmatic responses is high, i.e. NUMBER and SOME for about half of the participants of our experiments. The results of the baseline conditions in our experiments show that in these cases, strong primes do not boost the rate of pragmatic responses. Rather, the effect is driven by the inhibition effects of weak primes. To explain this fact, one could modify Bott & Chemla's explanation slightly and assume that in these cases, weak primes de-activated the relevant mechanism(s), and the de-activation had lingering effects in the target trial. However, we think this line of explanation still leaves some questions unanswered. Firstly, one would still expect effects of strong primes for the cases with high baseline rates, because forced activation should in principle be able to lead to an even higher rate of SI in the target trial. However, this is not what we observe in our results, and strong and weak primes generally have comparable effect sizes relative to the baseline rate. Secondly, the effect size of this type of priming is expected to be very small, given that the cross-scale priming effects observed by Bott & Chemla were generally very weak. Therefore, this alone is unlikely to explain the sizable inhibition effects of weak primes we observed for NUMBER and SOME for the relevant participants.

In order to explain the larger effect size for within-scale priming, i.e. priming with the same scalar expression, Bott & Chemla additionally propose that the alternatives used to generated SIs themselves can also be primed. The idea is that a strong prime

[10] Bott & Chemla (2016) discuss different possibilities for what aspects of the SI computation is actually primed, e.g., it could be the mechanism of searching for alternatives or the mechanism that generates inferences that amount to the negations of the alternatives, or both. These details are not important for the discussion here.

forces the use of a certain alternative in deriving a SI, and this alternative stays salient and active in the participant's mind when they subsequently do a target trial. Since the inverse preference effects we observed in our results are more robust than the cross-scale priming effects reported in Bott & Chemla (2016), this type of priming is more relevant for the discussion here.

However, we think that Bott & Chemla's hypothesis based on the activation level of alternatives themselves does not help us much in explaining the full inverse preference effects we observed in our results. The primary reason is because this type of priming cannot lead to the sizable inhibition effects of weak primes that we observed for cases with high baseline rates of pragmatic responses, i.e. Ad hoc and Some in the case of about half of our participants. In principle, the activation level of alternatives could be primed (although we claim later that we have no evidence for it in our results), but it would not make much sense to assume that not using an alternative and hence not deriving the corresponding SI in the weak primes would make this alternative stay non-salient in the target trial. If such an inhibition effect existed, no SIs should be very robust, because in reality the vast majority of occurrences of a scalar item are not immediately preceded by another occurrence, so the relevant alternative would be extremely de-activated and non-salient. Despite this, Number, for example, is preferentially interpreted with an SI, as its high baseline rate of pragmatic responses suggests. Therefore, we do not think the relatively large priming effects triggered by weak primes can be fully explained in terms of the activation level of the alternative itself.

The discussion so far certainly does not force one to renounce the idea of priming of the mechanism(s) of SI computation or priming of alternatives altogether, but it suggests that there is some other priming mechanism at play. To this end, we will suggest below that what is primed in the experimental paradigm under discussion is expectations about the type of conversational context one is likely to be in at a given moment, which indirectly affect the computation of SI. One major difference from the activation-based account is that priming is not directly caused by a high activation level of the mechanism behind SI computation and/or an alternative, but rather, it is the result of probabilistic reasoning about weather the SI is intended or not, which is done in a way that is affected by what happens in priming trials.

We furthermore claim that this type of priming is enough to explain the experimental results, and therefore the experimental paradigm in question might actually never directly prime the mechanism(s) of SI computation or alternatives *per se*, contrary to Bott & Chemla's claim.

6.2. Context adaptation

Besides SI, priming has also been employed to investigate a number of other linguistics phenomena, most notably structural priming of various syntactic constructions (e.g., the realisation of the goal argument of a di-transitive verb in English; Bock 1986; Hartsuiker & Kolk 1998; Hartsuiker & Westenberg 2000; Ferreira 2003; Kaschak et al. 2011; Fine et al. 2013; Jaeger & Snider 2013; see Pickering & Ferreira 2008 for an overview of syntactic and other structural priming in general, and Maldonado, Chemla & Spector 2017, 2019; Feiman & Snedeker 2016 for studies that use priming for other

semantic phenomena than priming). For these phenomena too, inverse preference effects have been observed, which generally poses an issue for an account that is based on salience and/or activation alone, as inhibition effects would be unexpected. To explain these inverse preference effects, Fine et al. (2013) and Jaeger & Snider (2013), among others, put forward an explanation based on the idea of *adaptation*. According to their view, priming effects arise from implicit learning and rapid online adaptation of probabilistic expectations about words and constructions that take place in the course of the experiment. We claim that implicature priming can and should be similarly understood in terms of adaptation, more specifically, adaptation of conversational context, and argue that it accounts for inverse preference effects as well as spillover effects.

SIs are known to be context-dependent in the sense that one and the same scalar item gives rise to an SI very robustly in some contexts, but most naturally receives the weak reading in other contexts. By way of illustration, let us consider *ad hoc* implicatures. For example, an utterance of *I have two daughters* is most likely understood with an *ad hoc* implicature that I don't have sons, if the utterance is meant to be an answer to the question *Do you have kids?*. But an utterance of the same sentence is likely to be read without this SI, if one intends to answer a different question like *Do you know anything about what small girls are interested in?*. Likewise, it is known that more canonical SIs, such as those of *some* and numerical expressions, are also context-dependent in a similar fashion. Concretely, in answering *What do you see on the card?*, *Some of the symbols are hearts* and *Four of the symbols are hearts* very robustly have SIs, which amount to 'not all of the symbols are hearts' and 'no more than four of the symbols are hearts', respectively. On the other hand, if someone says *I wonder if some of the symbols on your card are hearts*, the SI that not all of the symbols are hearts is at least cancellable, as in *Indeed, some of the symbols are hearts; in fact, all of them are*. Similarly for *four*: If someone says *I will lose if you have a card with four or more hearts*, one could say, *On my card, four of the symbols are hearts; in fact six of them are*.

By assumption, competent speakers of English are aware of the context-sensitivity of SIs, at least subconsciously. In actual conversations, furthermore, they sometimes find themselves in a situation where they do not have necessary information to determine with perfect certainty which reading is intended by the speaker. In such a case, it is reasonable to assume that they use whatever information available to them to make probabilistic hypotheses about which interpretation is intended.

Now recall our experimental setup, as in Bott & Chemla (2016), the participants were not instructed to imagine a particular conversational context, but perhaps they nonetheless made (probabilistic) assumptions about what kind of context they were likely to be in, when they interpreted the linguistic stimuli. For instance, they might have supposed that the linguistic stimuli were intended to answer some implicit question under discussion and/or to inform some hypothetical communicative agent with some specific goal.

For the baseline block of our experiments, there were not many useful cues for inferring what the intended context was like, other than the linguistic and visual stimuli themselves, and consequently it is likely that the participants made decisions based on their prior linguistic and other relevant experience and knowledge regarding each scalar expression. This could at least partially explain the variation in the baseline conditions

for the three scalar items we tested. That is, the participants assumed that Ad hoc is generally likely to be used in a context where no SI is intended, while Number is more likely to be used in a context where its SI is relevant. Furthermore, for Some, different participants had different ideas. About half of them thought it is more likely to be used in a context where its SI is intended, and the other half thought it can easily be used in a context where it is not intended. This still leaves open why such variation is only observed for Some, but we will leave this question open for now, and come back to it at the end of the paper.

Crucially, we claim that the priming effects in the experimental paradigm under discussion should be understood in terms of adaptation of such probabilistic expectations about the distributions of the weak and strong readings of a given scalar item. We assume that the mechanism responsible for forming relevant expectations is highly flexible, and can update the expectations rapidly and incrementally as more contextual cues come in. That such a flexible adaptation mechanism exists is not at all conceptually implausible, given that speakers in real speech contexts not only make assumptions about what kind of conversational contexts they are in, but also have to make such guesses and adjust their expectations constantly and rapidly, since the conversational context keeps growing and changing directions. Thus, for efficient linguistic communication to be possible in a constantly evolving conversation among agents, each of whom potentially only has partial information about it, it is useful to have a flexible mechanism that dynamically adjusts one's expectations according to relevant incoming cues, linguistic or not. There is independent evidence that expectations about various aspects of conversational context are indeed continuously and dynamically formed and updated during conversation (see Kuperberg & Jaeger 2016 for an overview and further references). Although, to the best of our knowledge, not much research has been conducted on SIs from this perspective, it would not be at all surprising if contextual expectations relevant for interpretations of scalar expressions could be analogously continuously and flexibly adapted.

When applied to implicature priming, the mechanism of context adaptation gives rise to priming effects in the following way. As explained above, we assume that for each scalar expression, the participant starts with some prior expectation about types of context it is likely to be used in, and this expectation manifests itself in the form of the probabilistic distribution of its weak and strong readings in the baseline conditions of our experiments. The conversational context normally does not change radically from one moment to another, at least without an explicit indication of change, so in the absence of such a cue they do not radically revise their prior expectations. However, in Block 2, they encounter weak and strong priming trials, and these can cause the participants to adapt their expectations. Specifically, strong primes indicate that the current context is one where the SI should be derived and weak primes indicate that the current context is one where the SI should not be derived. If one starts with a prior expectation that the scalar item in question is unlikely to be used in a context where its SI is intended, e.g., in the case of Ad hoc, then one will have to revise this prior expectation when one encounters a strong prime, because one can infer that at least in the context of this experiment, such otherwise surprising contexts are not at all unlikely, contrary to their initial expectation. Now, since the target item does not explicitly indicate that the context has changed radically, after priming trials, the participant is likely to keep

the same expectation that they had or formed during the priming trials, giving rise to a priming effect in the target trial. Scalar items for which the prior expectation is that they are used in a context where their SIs are intended e.g. NUMERAL, an analogous process explains priming effects on them. That is, in the course of the experiment, one encounters priming trials that forces one to revise one's prior expectation, and that has an effect on the following target trials.

This mechanism of context adaptation provides a natural explanation for the inverse preference effects for implicature priming. For AD HOC, for example, for which the prior expectation is that it is more likely to be used in a context where an SI is not intended, a weak prime would not have a lot of priming effect, because it would not require a substantial change in the prior expectation. On the other hand, a strong prime would demand a radical change in the expectation. Consequently, weak primes have very little priming effects, if any, but strong primes have larger priming effects. The effect will be the opposite for scalar items like NUMBER, for which the prior expectation is that they are more likely to be used in contexts where the SI is intended. That is, for such scalar items, weak primes have sizable priming effects, as they demand revising the prior expectation, while strong primes have very small effects, if any. Generally speaking, more surprising information results in a bigger adaptation effect, giving rise to a pattern that can be characterised as inverse preference.

The mechanism of context adaptation also provides a natural explanation of the spillover effects in our experimental results. According to the current account, participants adapted their expectations in Block 2, upon encountering priming trials that suggested to them that contexts that were unexpected according to their prior expectations could in fact be intended in this experiment. The following trials were then undertaken with these adapted expectations, and potentially fed further cycles of adaptation. If, at one point in the experiment, one had to perform a substantial revision of the current expectation, which should happen when one encounters a priming trial that demands a context type that is very surprising with respect to one's current expectation, then the following trials should show a residue of this adaptation, regardless of which experimental condition they are in, giving rise to spillover effects. Finally, how strong spillover effects should be is a function of long-lasting the adaptation effect should be.

In sum, according to our explanation, priming effects in this experimental paradigm arise through *context adaptation*: The participant starts the experiment with some prior expectations about what kind of context they are likely to be in for each scalar expression. Priming trials force one particular type of context to be used, providing cues about what kind of context one is currently in, and the participant adapts their expectation accordingly. Given that our experiment included both strong and weak primes for each scalar item and each participant had a preference for each scalar item with respect to the weak vs. strong readings, the participants inevitably encountered priming trials that forced an unexpected type of context for each scalar item, which led them to alter and adapt their prior expectations to whatever context or contexts they inferred this experiment was about.

We wish to make no strong theoretical commitments as to the exact nature of the context or context type that one reasons about in the experimental paradigm under discussion, as this question is largely open at this point. Above, we illustrated the context sensitivity of SIs in terms of so-called 'questions under discussion', but we would

like to leave it open whether such questions themselves can be equated with contexts about which one forms expectations about, or, alternatively, contexts should be seen as something more general that encompasses questions under discussion. Moreover, there are further analytical possibilities. For example, it could be that the relevant expectations are about some other contextual factors altogether that might or might not have directly to do with questions under discussion. More concretely, the expectations might be about the identity of the speaker, and the participants might have assumed that there were different speakers with different conversational goals behind the experiment (cf. Kleinschmidt & Jaeger 2015).[11] Further research is needed to settle questions like these.

7. Hidden Markov Models for context adaptation

In order to make the idea of context adaptation more concrete, we will present a computational implementation of it. Previous work on computational modelling of adaptation (Fine, Qian, Jaeger & Jacobs, 2010; Kleinschmidt, Fine & Jaeger, 2012; Fine et al., 2013, among others) has demonstrated that Bayesian belief-update models provide reasonable accounts of syntactic adaptation. In particular, it accounts for inverse preference effects thanks to the fact that more surprising information gives rise to larger learning effects. Concretely, all three studies on syntactic adaptation mentioned here use *beta-binomial* models that make probabilistic predictions about the distributions of different syntactic constructions. The parameters of these models get incrementally updated via Bayesian learning, as they encounter new data, and the models' predictions are adjusted accordingly.

However, as we will argue below, beta-binomial models are not flexible enough to capture our idea of context adaptation for implicature priming. For this reason, we will use a more general class of models called *Hidden Markov Models (HMMs)* (see, e.g., Ghahramani 2001; Rabiner & Juang 1986; Meeden & Vardeman 2000; Stamp 2018; Jurafsky & Martin 2020).

Recall that according to our account based on context adaptation, expectations relevant for implicature priming are not just about how likely the weak and strong readings of a given scalar item are observed in a given conversational context, but also about what kind of conversational context one is more likely to be in, given a linguistic stimulus. This context parameter is *hidden* from the perspective of the participants of our experiments in the sense that it cannot be directly observed and has to be inferred.

[11] Bott & Chemla (2016) critically discuss the possibility that the participants of their experiments postulated two (types of) speakers, one who intended the strong reading and one who intended the weak reading. They rejected this possibility, claiming that since there was no cover story, there was no principled reason for the participants to assume that there were multiple speakers, and that all items were presented one after another without a clear boundary that could indicate a change of speakers. We are not entirely convinced by their reasoning here. While the participants initially assumed that there was only one type of speaker, we think that priming trials that forced the less likely reading could have been good enough cues for them to reconsider that assumption. At this point, however, we do not have enough evidence for or against the idea that the participants reasoned about (hypothetical) speakers, and therefore would like to leave this as a theoretical possibility.

HMMs are useful in modelling reasoning and learning involving such hidden parameters. In particular, learning for HMMs amounts to simultaneously adapting (i) the expectations about the probabilistic distributions of different types of context and (ii) how likely each interpretation is to be observed in each type of context.

7.1. Basics of HMMs

More formally, a HMM operates on a sequence $\vec{o}$ of observations. In our case, each observation is either the strong reading (s) or the weak reading (w) of the scalar expression in question. Thus, $\vec{o}$ is a sequence of symbols drawn from the alphabet $O = \{s, w\}$. In order to simplify the discussion, we assume that they perfectly match their behavioral correlates in our experiment. That is, the participants chose the overt picture if and only if they understood the sentence under the weak reading.

Context type is the hidden parameter of our HMM. We assume that there are two possible types of context, + and −: + is the type of context that favors the strong reading and − is the type of context that favors the weak reading. The set of context types is denoted by $Q = \{+, -\}$. At a given moment, the model assumes one of these two context types. Also, at the next moment, it might or might not be assuming the same context type.

The model has some more components, which are the flexible bits that get updated as more information comes in.[12] Firstly, it has transition probabilities, $A = \{a_{++}, a_{+-}, a_{-+}, a_{--}\}$, which are probabilities of moving from one context type to another. Secondly, each context type c is associated with two emission probabilities, $b_c(s)$ and $b_c(w)$: $b_c(s)$ is the probability of observing s (the strong reading) in c, and $b_c(w)$ is the probability of observing w (the weak reading) in c. Finally, π is the initial probability distribution over states, specifying how likely it is that the HMM starts with each context.

Thus, the HMM we will be using below is characterised by a quintuple $\langle O, Q, A, B, \pi\rangle$, and the last three components are subject to update. The model structure is schematically depicted in Figure 8.

A HMM $\lambda = \langle O, Q, A, B, \pi\rangle$ makes probabilistic predictions as to what the next observation is. Or, equivalently for our purposes, the HMM can be seen as producing observations with some probabilities. Predictions for future observations can be computed using what is known as the *forward algorithm*. The forward algorithm recursively computes the probability $P(o_1 \cdots o_i, q_i = q|\lambda)$ of observing a particular sequence of observations $o_1 \cdots o_i$ of length i with the HMM ending in context type q at the last step i.

When $i = 1$, the probability is about the very first observation, meaning that the HMM has not done anything yet, and we are trying to predict which observation it produces. If the first context type that the HMM assumes happens to be +, then the probability that it will produce s will be $b_+(s)$, and if the first context type it assumes happens to be −, then the probability that it will produce s will be $b_-(s)$. We do not

[12] Strictly speaking, the alphabet and the hidden states could also be estimated and updated, but we fix them for the sake of simplicity. This assumption is not too farfetched in the case at hand, because we could assume that O and Q follow from one's linguistic knowledge.

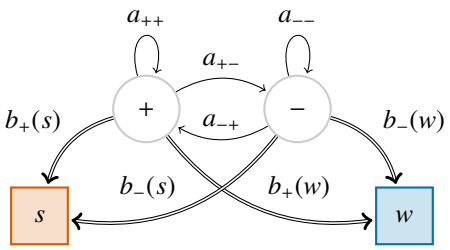


Figure 8: A schematic representation of the structure of the HMMs used to model the experimental results. + and − are context types, which are the hidden states of this HMM, and s and w are observations. The arcs are labelled with probabilities, which get updated via Bayesian learning. The initial probability distribution π is not depicted here.

know which context type it will assume in the first step, but we know how likely each of these scenarios is, namely, π_+ and π_-, respectively. Thus the probability of observing s will be the sum of $\pi_+ \cdot b_+(s)$ and $\pi_- \cdot b_-(s)$. More generally, for whichever observation o_1 and for each context type q:

$$P(o_1, q_1 = q|\lambda) = \pi_q \cdot b_q(o_q).$$

In computing the case of $i = 2$, the algorithm references the probability we have just computed by summing over different cases of q_1.

$$P(o_1 o_2, q_2 = q|\lambda) = \sum_{q' \in Q} P(o_1, q_1 = q'|\lambda) \cdot a_{q'q} \cdot b_q(o_2).$$

This second step is generalisable to any i as follows, making use of the *Markov property* of the HMM, i.e., the proper that its prediction depends only on the immediately preceding state.

$$P(o_1 \dots o_i, q_i = q|\lambda) = \sum_{q' \in Q} P(o_1 \dots o_{i-1}, q_{i-1} = q'|\lambda) \cdot a_{q'q} \cdot b_q(o_i).$$

Using the forward algorithm, the probability of observing a given sequence $\vec{o}$ can be computed by simply summing over the probabilities for different possible final contexts.

Conversely, when a particular sequence $\vec{o}$ of length t is given, it is also possible to compute the probability $P(q_i = q|\vec{o}, \lambda)$ of being in context $q \in Q$ at each step i for $1 \leq i \leq t$ of processing $\vec{o}$. This computation uses the forward algorithm as well as the so-called *backward algorithm*. The backward algorithm computes the probability $P(o_{i+1} \cdots o_t|q_i = q, \lambda)$ of observing the $t - i$ observations at the tail of $\vec{o}$, assuming that the context is $q \in Q$ at step i. To illustrate, if $i = t$, there's no more data, so for any context $q \in Q$, this probability is simply 1. If $i = t - 1$, then we can obtain this probability by summing over the probability of observing o_t for different prior context at step $t - 1$:

$$P(o_t|q_{t-1} = q, \lambda) = \sum_{q' \in Q} a_{q'q} \cdot b_{q'}(o_t).$$

Using this probability, we can compute the probability $P(o_{t-1}o_t|q_{t-2} = q, \lambda)$ for $i = t-2$ as

$$P(o_{t_1}o_t|q_{t-2} = q, \lambda) = \sum_{q' \in Q} a_{q'q} \cdot b_{q'}(o_t) \cdot P(o_t|q_{t-1} = q, \lambda).$$

It is easy to see that this can be generalised for any i such that $1 \leq i \leq t$.

Now, with the forward and backward algorithms, we can compute the probability $P(q_i = q|\vec{o}, \lambda)$ of the HMM λ being in context $q \in Q$ at step i of processing $\vec{o}$ as

$$P(q_i = q|\vec{o}, \lambda) \propto P(o_1 \cdots o_i, q_i = q|\lambda) \cdot P(o_{i+1} \cdots o_t|q_i = q, \lambda)$$

7.2. Prior models

We will use the above HMM to model the results of Experiment 1. In particular, we would like to see if it successfully predicts inverse preference and spillover effects.

The model will be updated via Bayesian learning as it processes data, but we need to start with some initial model. In this case, however, it is not easy to make empirical estimates of realistic initial settings, especially because there is a hidden contextual parameter that we cannot observe directly. We thus start with the following somewhat contrived parameter settings that are skewed towards the intended interpretation such that the strong reading (s) is more likely in context +, the weak reading (w) is more likely in context −, and it is more likely to stay in the same context type.

$$a_{++} = 0.75, a_{+-} = 0.25, a_{-+} = 0.25, a_{--} = 0.75$$
$$b_+(s) = 0.75, b_+(w) = 0.25, b_-(s) = 0.25, b_-(w) = 0.75$$
$$\pi(+) = 0.5, \pi(-) = 0.5$$

Before modelling the priming data, we train this initial model using the empirical data from the baseline block to create models of different participants at the stage before they did the priming block of our experiment. Specifically, for each scalar item s and for each participant a of Experiment 1, we took the proportion $p_{s,a}$ of 'Better Picture?' answers in the Baseline condition, and trained the model with a random sequence of observations drawn from the Bernoulli distribution whose expected value is p, i.e., bernoulli($p_{s,a}$). We also appended a sequence ws at the beginning of the training sequence, in order to avoid a uniform sequence, which often leads to extreme parameter values (cf. Rabiner & Juang, 1986).

The length of the training data is a free parameter here, and controls how much effect the training will have, which means how faithful the resulting model will be to the participant's actual behavior in the baseline block. But there is a trade-off: It will also affect how flexible the model will be in the priming block. That is, the longer the training sequence, the smaller the effects of new observations will be, so the model will be less flexible in the priming block. We picked the length of training sequences that maximises the data likelihood in the priming phase, which happens to be 10 including the two fake observations at the beginning.

Learning for our model amounts to Bayesian belief update via Maximal Likelihood estimation. That is, given a sequence of observations $\vec{o}$, the model is updated with the best parameter values $\langle A, B, \pi \rangle$ that maximise the data likelihood $P(\vec{o}|A, B, \pi)$. For a

HMM, however, the cost of directly computing data likelihood grows exponentially as a function of data length. Luckily, there is a solution. It is common to estimate the best parameter values using the *forward-backward algorithm* (a.k.a. Baum-Welch algorithm), which is a kind of Expectation Maximisation algorithm that reduces the computational cost via Dynamic Programming. It is so-called because it makes use of both forward and backward algorithms, which we introduced above, but we refrain from discussing details of this procedure here to save space, and refer the interested reader to the introductory works cited at the beginning of this section. In addition, our code for the forward-backward algorithm applied to our HMMs, which is based on the pseudo-code in Stamp (2018, Ch. 2), is available at https://osf.io/263xf/.

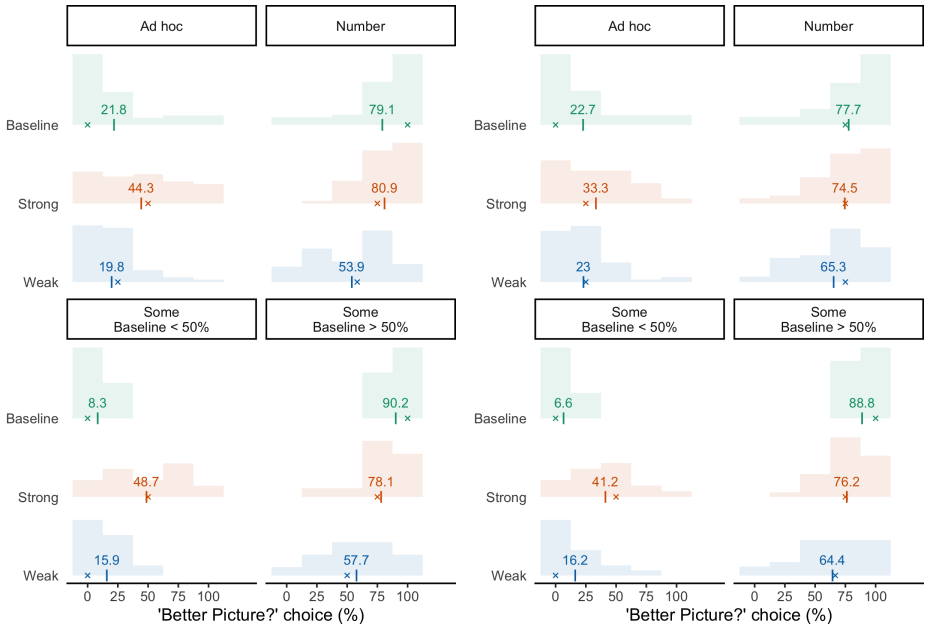
7.3. Modelling the priming effects

The training procedure explained above resulted in three baseline models for each participant, one for each of the three scalar expressions we tested in our experiments. We now model the priming data.[13]

The baseline HMM models we obtained by training the prior model with the results of the baseline block are used to model the priming data as follows. Each baseline model, which is meant to be a model of a particular participant's behavior with respect to a particular scalar expression based on their prior expectations, encounters the priming trials in the order that the participant that this model is modelling actually encountered them in Experiment 1, and the answers that the participant actually gave to these priming trials were treated as new observations for the model. Specifically, for each target trial that were preceded by two priming trials to which the participant gave the correct answers, we updated the model with the answers to the two priming trials, and computed the probability p of observing s with respect to the updated model. Then this probability was used to make a choice as to which reading to get in the target trial, by picking a random observation from the Bernoulli distribution with an expected value of p, bernoulli(p). In case the two priming trials preceding the target trial were not answered correctly, we did not compute the model prediction, which corresponds to the omission of such results in our data analysis for Experiments 1-3. The priming trials that were given incorrect answers were also simply ignored and were not used to update the model, as we do not know which reading the participant had in mind when they gave these incorrect answers.

The results of the model predictions are summarised in Figure 9a. The Baseline conditions generally mirror the actual results, as expected. Importantly, furthermore, we observe inverse preference patterns here. Specifically, for Ad hoc, the baseline rate

[13]Note that we are modelling the three scalar expressions separately, which essentially means that we ignore cross-scale prime effects altogether. As Bott & Chemla (2016) and Meyer & Feiman (2021) have observed, however, cross-scale priming does have an effect, and it is reasonable to assume that part of our experimental data should also be understood in terms of cross-scale priming and its spillover effects. However, at the same time, these previous studies also observed that the effect size of cross-scale priming is generally small, and so we don't expect large spillover effects. We therefore believe that our negligence of cross-scale priming here does not undermine the results reported below and our main conclusions based on them. It should also be emphasised that there is nothing incompatible between context adaptation and cross-scale priming. We will come back to the question of how to understand cross-scale priming in terms of context adaptation in the next section.



(a) HMMs

(b) Beta-binomial models

Figure 9: The predictions of (a) the HMMs and (b) beta-binomial models trained by the data from Experiment 1, as explained in the text. For each condition, the distribution of by-participant mean proportions is visualised by a histogram, the grand mean by a thick bar with its value on top, and the median by a cross.

is low, and strong primes had boosting effects, while weak primes had no inhibition effects. For NUMBER, on the other hand, we see inhibition effects of weak primes, but no boosting effects of strong primes are observed.

For SOME, the baseline models replicate the bi-modal distribution in the Baseline block, and the results are already broken down in Figure 9a according to the baseline rate, as we did for the actual experimental results. These predictions closely resemble the actual results as summarised in Figure 4. In particular, we see clear inverse preference effects for both groups.

It should also be remarked that spillover effects are observed in the model predictions. Recall that in the actual results, a spillover effect of weak priming was clearly seen in NUMBER in Experiment 1. A similar effect is observed between the medians for BASELINE and STRONG in Figure 9a. Similarly, in the actual results, the breakdown of the results for SOME by the baseline preference clarified the spillover effects. This is very nicely replicated in the modelling results.

These spillover effects are in fact expected, given the architecture of the HMM. That is, the training based on the results of the Baseline block created a bias in the model one way or another, and in the priming block, it encountered observations that were unexpected given this bias, and readjusted the parameter values so that these data points became less surprising. This readjustment affected the model predictions in later trials regardless of the experimental condition, leading to spillover.

7.4. Beta-Binomial models

We have just seen that HMMs embodying our account based on context adaptation nicely capture the main observations, but in order to further strengthen this conclusion, we will now compare the above predictions of HMMs with those of beta-binomial models. As mentioned at the beginning of this section, previous studies on syntactic adaptation used beta-binomial models. Beta-binomial models can seen as a special class of HMMs whose hidden parameters are trivial. Concretely, the beta-binomial version of our HMM will be a model whose context-type parameter is fixed, and whose predictions are just about how likely *s* and *w* are.

The comparison between HMMs and beta-binomial models is also of interest, given the previous studies on syntactic adaptation that used beta-binomial models, but more importantly, since beta-binomial models cannot reason about the hidden context-type parameter, if their performance is worse than that of HMMs, that will give us further support for the hypothesis that implicature priming has to do with adaptation of a hidden parameter, context type.

The beta-binomial models we will use were constructed in the same way as the HMMs above except that one of the context types is removed and the initial parameter values were re-adjusted accordingly. These models were then trained in exactly the same manner using the actual results of Experiment 1. As before, the code we used to construct our beta-binomial models is available online at https://osf.io/263xf/.

The predictions of these beta-binomial models are summarised in Figure 9b. As these graphs show, the differences between the two priming conditions are generally attenuated in comparison to the predictions of HMMs. This makes sense because the beta-binomial models try to guess the overall distributions of the strong and weak readings in probabilistic terms, and in the priming block of our experiments, they are equally likely from a global point of view, although locally, i.e., at each point in the experiment, they were not completely equally likely. In other words, what these models try to learn is the fact that each reading is roughly 50% likely in the current experiment, and in fact the predictions for the priming conditions are generally closer to 50% than the baseline rates. We nonetheless observe slight priming effects in the same direction as before. In particular, the breakdown of the predictions for Some according to the baseline preference, as summarised in Figure 9b, looks generally reasonable. However, these priming effects are simply due to the combined effect of more surprising information having larger learning effects, which gives rise to an inverse preference pattern, and the fact that at the point of a given target trial, the distributions of the weak and strong readings up to that point are not exactly 50-50. In fact, if we had more priming trials, the differences between strong and weak priming would be expected to balance out. We can say, therefore, the inverse preference effects are not captured very well by the beta-binomial models.

On the other hand, there are massive spillover effects in the sense that the two priming conditions overall look alike, as the histograms for these conditions suggest. This again is as expected, because what the beta-binomial models do is guess the overall rates of the two readings, and in the current experiment, the two readings are equally likely. Consequently, the beta-binomial models predict too much spillover.

Note that HMMs are generally more flexible, as they have more structure and parameters to readjust, so it is not very surprising that HMMs make better empirical pre-

	AIC	BIC
HMMs	1348.809	1374.184
Beta-binomial models	1418.213	1423.288

Table 5: The AIC and BIC values of the HMMs and beta-binomial models with respect to the results of the priming block of Experiment 1. Smaller scores indicate better models.

dictions. In order to numerically compare the two classes of models while taking this difference in flexibility into account, we use the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), which are standardly used for model comparison. As shown in Table 5, the HMMs perform better than the beta-binomial models with respect to both AIC and BIC. This comparison gives further support to our hypothesis that the adaptation mechanism used in the experimental paradigm under discussion involves an inference about a hidden parameter.

Before moving on, it should be noted that more complex versions of HMMs might be able to explain the data even beetter. For instance, we assumed that our HMMs reason about two possible context types, but theoretically there could well be more context types with different effects on SIs. Whether such a more complex model is desirable is essentially a theoretical question, as how feasible it is as a model of human reasoning needs to be evaluated against a theory of context adaptation. We must admit that we do not have precise enough theoretical ideas about the notion of context type at this moment, and consequently, our conclusion here needs to stay modest: The comparison between the beta-binomial models and the simple HMMs suggests that implicature priming involves reasoning about a hidden parameter, but the exact nature of the hidden parameter remains as an open question.

8. General Discussion and Conclusion

Our experiments were designed to clarify the direction of priming effects with the help of a baseline block. The baseline block was placed before any priming was introduced, providing neutral baseline results that are free from potential priming effects coming from other experimental trials within the same block, unlike the baseline trials of previous studies (Bott & Chemla, 2016; Waldon & Degen, 2020). Our baseline results clearly show that priming effects in the current experimental paradigm are best characterised as inverse preference: Priming trials that force the reading that is disfavoured in the baseline have large priming effects, and priming trials that force the reading that is preferred in the baseline have small or no priming effects. Furthermore, our experimental results and the comparison between the results of Experiments 1–2 suggest that large priming effects can spill over to other trials in the same block, justifying our concern about the baseline conditions of previous studies.

These two observations teach us important methodological lessons about the priming paradigm: Neutral baselines are necessary to know the direction of priming effects, and mixing different experimental conditions in a single block of trials can lead to unintended spillover effects that could make the results harder to interpret.

In addition, our experimental results have theoretical implications. In particular, we argued that accounts of implicature priming based on boosting the activation level of a SI computation mechanism and/or of an alternative itself needs to be reconsidered, because they cannot capture the sizable inhibition effect of weak priming observed in cases when the baseline rate of pragmatic responses is high.

We offered an alternative account based on the idea of context adaptation, according to which implicature priming involves adaptation of probabilistic expectations about what the current conversational context is meant to be. In particular, priming effects arise by updating such expectations rapidly and continuously during the experiment, by using priming trials as cues for what the current context is intended to be like. We implemented this idea using HMMs, which demonstrated that context adaptation via Bayesian learning can account for inverse preference, as well as spillover effects.

We would like to stress at this point that we think that context adaptation alone is enough to explain the key observations from this and previous studies. Note that the mechanism of context adaption could potentially co-exist with other mechanisms of priming. Recall, in particular, that Bott & Chemla (2016) postulated two different kinds of priming effects, priming effects coming from strong primes due to an increased activation level of alternatives, and priming effects due to an increased activation level of the mechanism for searching for or negating alternatives. Context adaptation is theoretically compatible with both of these, both individually and simultaneously, but we argue below that there is no strong evidence that such multiple routes to priming effects co-exist in our experimental results.

Firstly, we claim that it is unlikely that the activation level of alternatives referenced in strong primes has boosting priming effects in implicature priming, contrary to Bott & Chemla (2016) (see also Rees & Bott, 2018; Waldon & Degen, 2020, for related claims). This is because if it were relevant, then strong primes should generally have larger priming effects than weak primes, given the same level of prior expectation. To be concrete, let us hypothetically suppose that for Ad hoc, the baseline rate of pragmatic responses is 25% and for Number, the baseline rate of pragmatic responses is 75%. Then, context adaptation predicts that for Ad hoc, strong primes should have larger adaptation effects than weak primes, and conversely for Number, weak primes should have larger adaptation effects than strong primes. Furthermore, it should be that strong primes for Ad hoc and weak primes for Number should have comparable adaptation effects, given that they are equally surprising according to the prior expectations. Now, if the activation level of alternatives could also lead to priming effects, then we would expect an asymmetry between weak and strong primes such that strong primes should have larger overall priming effects, because their priming effects are combined effects of context adaptation and activated alternatives. On the other hand, the priming effects of weak primes only involve context adaptation. In our experimental results, however, we do not observe such an asymmetry between strong and weak priming. In fact, if anything, the weak primes tended to have larger effects in both Experiment 1 and Experiment 2.

The above discussion casts doubt on the idea that the activation level of an alternative is involved in implicature priming, but it should also be noted that this is a particularly attractive idea, given that there is independent evidence in the literature suggesting that that it is a factor that can affect SI computation (e.g., Barner, Brooks

& Bale, 2011). Furthermore, there are several previous studies on implicature priming whose results also are taken to support this hypothesis, namely Rees & Bott (2018); Waldon & Degen (2020) and Marty, Romoli, Sudo & Breheny (2021). These studies tested what is called *alternative priming*, in addition to strong and weak priming. In alternative priming, the relevant alternative is presented in the priming trials preceding the target trial. In the case of SOME, for example, one will see sentences involving *all* in the priming trials, and then in the target trial, one will see a sentence with *some*. The results reported in the three studies all suggest that alternative priming indeed has robust boosting effects, at least for some scalar items (see below). This can be taken as evidence that increased activation level of an alternative can indeed have priming effects, and if that is correct, then we will have conflicting evidence for the role of the saliency of alternatives for implicature priming.

We would like to suggest ways to resolve this conflict. The first possibility is that there might be a difference between an overt presentation of an alternative, as in alternative priming, and a covert reference to an alternative, as hypothetically involved in strong priming in the experimental paradigm under discussion. It seems reasonable to assume that the former has a larger effect than the latter on salience. Although this should of course be empirically verified, assuming it is on the right track, we could explain the difference between alternative and strong priming with respect to salience, by assuming that the salience level increased by strong primes is not high enough to have an effect on the target trial, while the salience level increased by alternative primes is very high and it manifest itself in the target trial as a priming effect.

Another, more radical, theoretical possibility is that alternative priming should in fact be understood in terms of context adaptation as well, without recourse to salience or activation. To illustrate this idea concretely, let us consider alternative priming for SOME again. The alternative primes involve *all*, and from these trials one could potentially learn something about the current conversational context. These alternative primes suggest that *all* is a relevant expression in the current context, and that could affect one's expectations about what type of context is intended in the subsequent target trial involving *some*. If this second explanation is on the right track, then the increased salience of alternatives might not be relevant at all for implicature priming.

The results of our experiments do not provide convincing evidence to tease apart these two possible accounts, and for this reason, we leave it to further research shed light on this issue. However, given these considerations, it is possible that an activation-based priming mechanism may have no role to play in implicature priming tested in the current experimental paradigm, contrary to what has been assumed since Bott & Chemla (2016).

Now, how about the other mechanism that Bott & Chemla (2016) claim can be primed? Recall that they took cross-scale priming as evidence for a common mechanism involved in the generation of the SIs for AD HOC, SOME, and NUMBER, and explain it in terms of this mechanism being activated in strong primes and staying active in subsequent trials (see also Meyer & Feiman 2021). This type of priming effect is generally very small, given the small effect size of cross-scale priming in the results. We do not have direct empirical evidence for or against the involvement of this mechanism at this moment, but we would like to remark that context adaptation alone is in principle enough to explain the results of our experiment, because context adaptation can

also explain the cross-scale priming effects observed in the previous studies mentioned here. To illustrate the idea, let us consider the case of strong vs. weak primes involving SOME priming NUMBER. In the results of Experiment 3 reported in Bott & Chemla (2016), more pragmatic responses were observed after strong primes than after weak primes, although the effect size was very small, namely 64.2% vs. 56.8%. Due to the lack of a neutral baseline in their design, we cannot know for sure which type of priming is responsible for the difference but they estimate the baseline rate to be 59% based on other trials (see fn. 5). Let us therefore assume for the sake of argument that strong primes involving SOME had booting effects on NUMBER. Context adaptation can provide a possible explanation of this observation without recourse to the mechanism of negating alternatives as follows. Upon encountering strong primes involving SOME, the participant learns that in the current context its SI implicature is intended, which can be taken to indicate that the speaker cares whether all the symbols on the relevant card are identical or not. Now in the target trials, the context might or might not stay the same, especially given that the linguistic stimulus looks different from the priming trials, but let us say that in the participant's probabilistic reasoning, the strong primes have increased the probability that the context is still the one where the speaker is likely to care whether all the symbols are identical or not. Then, they are more inclined to read the numeral *four* as 'exactly four'.

Admittedly this idea still needs to be worked out in more detail, especially with respect to what exactly the relevant notion of context consists in (see below for some possibilities), but we would like to also note that the small effect size commonly observed for cross-scale priming naturally falls from this account: Context adaptation has larger effects when the same expression or construction is involved, because one can more easily generalise what one learns in the priming trials to a target trial that superficially looks similar, than to a target trial involving a different linguistic stimulus, because by hypothesis the relevant contextual inferences are made based on the linguistic and visual stimuli for each trial. This possibility has a further theoretical implication: Cross-scale priming does not convincingly show that there is a common mechanism for different scalar inferences. In particular, the proposed account based on context adaptation is neutral with respect to how the two readings of a given scalar item are derived, and therefore could account for the cross-scale priming effects between NUMBER and SOME, even if it turned out that the relevant inference of NUMBER was not an SI, as proposed by some (cf. Geurts, 2006; Breheny, 2008), as long as the inference somehow affects one's expectations about the current context of utterance.[14]

In light of the above discussion, we would like to conjecture here that context adaptation is the only mechanism that is needed to explain implicature priming in the current

[14]Note that not everything seems to give rise to cross-scale priming effects on NUMBER and SOME. In Experiment 3 of Bott & Chemla (2016) for example, what they call PLURALS priming trials seem to have had no priming effects on SOME and NUMBER, and similarly Meyer & Feiman (2021) report results suggesting that FREE CHOICE has no priming effects on NUMBER and SOME, although the lack of neutral baselines in these studies make it hard to know for sure if there was no priming effects, as what is reported is simply that weak and strong priming resulted in the same rate of hidden card choices. In our view, these observations can be understood as resulting from these priming trials failing to affect aspects of context that matter for the scalar inferences of SOME and NUMBER, but we need to leave it open for now what these aspects exactly are.

experimental paradigm, contrary to Bott & Chemla (2016). However, a lot more needs to be understood about the mecanism of context adaptation. In particular, it needs to be made clearer what exactly the relevant expectations are about. We have hinted at the possibility that the question under discussion is a relevant aspect of context, as it is known to affect the robustness of SIs, but it is far from clear if the participants of our experiments actually reasoned about potential questions under discussion. Another potentially relevant aspect of context is the identity of the speaker. As mentioned before, context adaptation might involve reasoning about who the speaker is and what their conversational goal is. Note also that in theory, the question under discussion and the speaker's identity can be reasoned about simultaneously. In fact, there can be other aspects of context that might be affected by the priming trials of our experiments. Since the context is 'hidden' from observation by assumption, our experimental results do not provide direct evidence about what the participants actually reasoned about. Thus, to gain further insights on this question, we will need other experimental techniques, and this question is left open here for now.

In addition, the current study points to a couple of other directions for further research that should be mentioned here. Firstly, as we already remarked, we borrowed the idea of context adaptation from previous work on syntactic priming (Fine et al., 2013; Jaeger & Snider, 2013; Kuperberg & Jaeger, 2016). It is currently an open question if implicature priming and syntactic priming are to be explained by the same mechanism. If the answer to this question is yes, then it further supports our conjecture above, according to which this experimental paradigm does not directly affect the mechanism of SI generation. Secondly, we observed variation among scalar items with respect to their baseline rates. According to our account based on context adaptation, the variation arises due to speakers' prior linguistic experience with them. For instance, speakers learn through experience that Ad hoc is overall less likely to be used in a context where the relevant SI is intended. However, the variation we observed for Some poses an interesting question. Certainly, people should have different linguistic experience, but that should equally apply to all scalar items, and in general we do not expect speakers to have radically different experiences with respect to a extremely high frequent word like *some*. Thus, we think the inter-speaker variation we observed for Some might be due to some other factor. For instance, it is known that Some can be pronounced with or without a focal accent on it, and that seems to have an effect on how robust the SI is perceived to be (cf. Degen, 2015). The inter-speaker variation on Some could be due to different ways of reading the linguistic stimuli, which we and the previous studies on implicature priming did not control for.

Acknowledgements

This research was supported by the Leverhulme Trust grant RPG-2018-425.

Appendix A. Instructions for Experiments 1 and 2

General – In this study, we will ask for your judgments about English sentences. Every sentence that you will see will be accompanied by two pictures. Your task is

to decide which of the two pictures you think the sentence is describing. The study has two parts, Part 1 and Part 2, which slightly differ from one another. Please read carefully the instructions provided to you before you start each part.

Part 1 – Every sentence will be accompanied by two pictures: one of them will be visible to you, while the other one will remain covered with the label 'Better picture?' on it. The sentence is meant to describe **one and only one of these two pictures**. Your task is to decide which picture you think the sentence is describing: the visible one or the covered one? You will click on the visible picture if you consider it a match for the sentence; otherwise, you will click on the covered picture.

Part 2 – As in Part 1, every sentence will be accompanied by two pictures. In some cases, one of them will remain covered just as before but, in others, both pictures will be visible to you. As before, the sentence is meant to describe **one and only one of these two pictures** and your task is to decide which picture you think the sentence is describing. You will click on the picture that you consider a better match for the sentence.

Appendix B. Instructions for Experiment 3

General – In this study, we will ask for your judgments about English sentences. Every sentence that you will see will be accompanied by two pictures. Your task is to decide which of the two pictures you think the sentence is describing. The study has two parts, Part 1 and Part 2, which slightly differ from one another. Please read carefully the instructions provided to you before you start each part.

Part 1 – Every sentence will be accompanied by two pictures: one of them will be visible to you, while the other one will remain covered with the label 'Better picture?' on it. The sentence is meant to describe **one and only one of these two pictures**. Your task is to decide which picture you think the sentence is describing: the visible one or the covered one? You will click on the visible picture if you consider it a match for the sentence; otherwise, you will click on the covered picture.

Part 2 – The task of Part 2 is the same as Part 1, except that you will sometimes see pages with a single picture and no sentence. When you see such a page, please click on the displayed picture to proceed.

References

Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition*, *118*, 84–93. doi:10.1016/j.cognition.2010.10.010.

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*, .

Bates, D., Maechler, M., & Bolker, B. (2011). lme4: Linear mixed-effects models using s4 classes.

Bergen, L., Levy, R., & Goodman, N. (2016). Pragmatic reasoning through semantic inference. *Semantics & Pragmatics*, *9*, 1–83. doi:10.3765/sp.9.20.

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*, 355–387. doi:10.1016/0010-0285(86)90004-6.

Bock, K., Dell, G. S., Chang, F., & Onishi, K. H. (2007). Persistent structural priming from language comprehension to language production. *Cognition*, *104*, 437–458. doi:10.1016/j.cognition.2006.07.003.

Bock, K., & Griffin, Z. M. (2000). The persistence of structural priming: Transient activation or implicit learning? *Journal of Experimental Psychology: General*, *129*, 177–192. doi:0.1037/0096-3445.129.2.177.

Bott, L., & Chemla, E. (2016). Shared and distinct mechanisms in deriving linguistic enrichment. *Journal of Memory and Language*, *91*, 117–140.

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language*, *51*, 437–457. doi:10.1016/j.jml.2004.05.006.

Branigan, H. P., Pickering, M. J., Steward, A. J., & Mclean, J. F. (2000). Syntactic priming in spoken production: Linguistic and temporal interference. *Memory and Cognition*, *28*, 1297–1302. doi:10.3758/BF03211830.

Breheny, R. (2008). A new look at the semantics and pragmatics of numerically quantified noun phrases. *Journal of Semantics*, *25*, 93–139. doi:10.1093/jos/ffm016.

Chemla, E., & Singh, R. (2014). Remarks on the experimental turn in the study of scalar implicature, part i. *Language and Linguistics Compass*, *8*, 373–386. doi:10.1111/lnc3.12081.

Chierchia, G., Fox, D., & Spector, B. (2012). Scalar implicature as a grammatical phenomenon. In C. Maienborn, K. von Heusinger, & P. Portner (Eds.), *Semantics: An International Handbook of Natural Language Meaning* (pp. 2297–2331). Berlin: de Gruyter.

Degen, J. (2015). Investigating the distribution of *some* (but not *all*) implicatures using corpora and web-based methods. *Semantics & Pragmatics*, *8*, 1–55. doi:10.3765/sp.8.11.

Feiman, R., & Snedeker, J. (2016). The logic in language: How all quantifiers are alike, but each quantifier is different. *Cognitive psychology*, *87*, 29–52.

Ferreira, V. S. (2003). The persistence of optional complementizer production: Why saying "that" is not saying "that" at all. *Journal of Memory and Language*, *48*, 379–398. doi:10.1016/S0749-596X(02)00523-5.

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Ting, Q. (2013). Rapid expectation adaptation during syntactic comprehension. *PLOS One*, *8*, e77661. doi:10.1371/journal.pone.0077661.

Fine, A. B., Qian, T., Jaeger, T. F., & Jacobs, R. A. (2010). Is there syntactic adaptation in language comprehension? In *CMCL '10: Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics* (pp. 18–26).

Foppolo, F., Guasti, M. T., & Chierchia, G. (2012). Scalar implicatures in child language: Give children a chance. *Language learning and development*, *8*, 365–394.

Foppolo, F., Mazzaggio, G., Panzeri, F., & Surian, L. (2021). Scalar and ad-hoc pragmatic inferences in children: guess which one is easier. *Journal of Child Language*, *48*, 350–372.

Geurts, B. (2006). Take 'five': The meaning and use of a number word. In *Non-definiteness and plurality*. John Benjamins.

Geurts, B. (2010). *Quantitiy Implicatures*. Cambridge: Cambridge University Press.

Ghahramani, Z. (2001). An introduction to Hidden Markov Models and Bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, *15*, 9–42.

Grice, P. H. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Hartigan, J. A., & Hartigan, P. M. (1985). The dip test of unimodality. *The annals of Statistics*, (pp. 70–84).

Hartsuiker, R. J., & Kolk, H. H. J. (1998). Syntactic persistence in Dutch. *Language and Speech*, *41*, 143–184. doi:10.1177/002383099804100202.

Hartsuiker, R. J., & Westenberg, C. (2000). Word order priming in written and spoken sentence production. *Cognition*, *75*, B27–B39. doi:10.1016/S0010-0277(99)00080-3.

Hirschberg, J. B. (1991). *A Theory of Scalar Implicature*. New York: Garland.

Horowitz, A. C., Schneider, R. M., & Frank, M. C. (2018). The trouble with quantifiers: exploring children's deficits in scalar implicature. *Child development*, *89*, e572–e593.

Hunt, L., Politzer-Ahles, S., Gibson, L., Minai, U., & Fiorentino, R. (2013). Pragmatic inferences modulate n400 during sentence comprehension: Evidence from picture–sentence verification. *Neuroscience Letters*, *534*, 246–251.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of memory and language*, *59*, 434–446.

Jaeger, T. F., & Snider, N. E. (2013). Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, *127*, 57–83. doi:10.1016/j.cognition.2012.10.013.

Jurafsky, D., & Martin, J. H. (2020). Speech and language processing. URL: https://web.stanford.edu/~jurafsky/slp3/ draft.

Kaschak, M. P. (2007). Long-term structural priming affects subsequent patterns of language production. *Memory and Cognition*, *35*, 925–937. doi:10.3758/BF03193466.

Kaschak, M. P., & Borreggine, K. L. (2008). Is long-term structural priming affected by patterns of experience with individual verbs? *Journal of Memory and Language*, *58*, 862–878. doi:10.1016/j.jml.2006.12.002.

Kaschak, M. P., Kutta, T. J., & Jones, J. L. (2011). Structural priming as implicit learning: Cumulative priming effects and individual differences. *Psychonomic Bulletin & Review*, *18*, 1133–1139. doi:10.3758/s13423-011-0157-y.

Kaschak, M. P., Loney, R. A., & Borreggine, K. L. (2006). Recent experience affects the strength of structural priming. *cognition*, *99*, B73–B82. doi:10.1016/j.cognition.2005.07.002.

Katsos, N., & Cummins, C. (2010). Pragmatics: From theory to experiment and back again. *Language and Linguistics Compass*, *4*, 282–295.

Kleinschmidt, D. F., Fine, A. B., & Jaeger, T. F. (2012). A belief-updating model of adaptation and cue combination in syntactic comprehension. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 34. URL: https://escholarship.org/uc/item/6qm5v571.

Kleinschmidt, D. F., & Jaeger, F. T. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, (pp. 148–203). doi:10.1037/a0038695.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*, 32–59. doi:10.1080/23273798.2015.1102299.

Maechler, M. (2013). Package 'diptest'. *R Package Version 0.75–5*, .

Maldonado, M., Chemla, E., & Spector, B. (2017). Priming plural ambiguities. *Journal of Memory and Language*, *95*, 89–101. doi:10.1016/j.jml.2017.02.

Maldonado, M., Chemla, E., & Spector, B. (2019). Revealing abstract semantic mechanisms through priming: The distributive/ collective contrast. *Cognition*, *182*, 171–176. doi:10.1016/j.cognition.2018.09.009.

Marty, P., Romoli, J., Sudo, Y., & Breheny, R. (2021). Distinguishing between alternatives. Online poster presentation at Semantics and Linguistic Theory (SALT) 31.

Meeden, G., & Vardeman, S. (2000). A simple hidden Markov model for Bayesian modeling with time dependent data. *Communications in Statistics - Theory and Methods*, *29*, 1801–1826. doi:10.1080/03610920008832579.

Meyer, M.-C., & Feiman, R. (2021). Priming reveals similarities and differences between three purported cases of implicature: Some, number and free choice disjunctions. *Journal of Memory and Language*, *120*. doi:10.1016/j.jml.2020.104206.

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

Noveck, I. A. (2018). *Experimental Pragmatics*. Cambridge: Cambridge University Press.

Noveck, I. A., & Posada, A. (2003). Characterizing the time course of an implicature: An evoked potentials study. *Brain and language*, *85*, 203–210.

Pickering, M. J., & Ferreira, V. S. (2008). Structural priming: A critical review. *Psychological Bulletin*, *134*, 427–459.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: https://www.R-project.org/.

Rabiner, L. R., & Juang, B.-H. (1986). An introduction to Hidden Markov Models. *IEEE ASSP Magazine*, *3*, 4–16. doi:10.1109/MASSP.1986.1165342.

Raffray, C. N., & Pickering, M. J. (2010). How do people construct logical form during language comprehension? *Psychological science*, *21*, 1090–1097.

Rees, A., & Bott, L. (2018). The role of alternative salience in the derivation of scalar implicatures. *Cognition*, *176*, 1–14.

Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy*, *27*, 367–391. doi:10.1023/B:LING.0000023378.71748.db.

Spector, B. (2013). Bare numerals and scalar implicatures. *Language and Linguistics Compass*, *7*, 273–294. doi:10.1111/lnc3.12018.

Stamp, M. (2018). *Introduction to Machine Learning with Applications in Information Security*. Boca Raton, FL: CRC Press. doi:10.1201/9781315213262.

Teresa Guasti, M., Chierchia, G., Crain, S., Foppolo, F., Gualmini, A., & Meroni, L. (2005). Why children and adults sometimes (but not always) compute implicatures. *Language and cognitive processes*, *20*, 667–696.

Waldon, B., & Degen, J. (2020). Symmetric alternatives and semantic uncertainty modulate scalar inference. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*.