# Probability and processing speed of scalar inferences is context-dependent

**names**
{names}@school
Address

## Abstract

Studies addressing the question of whether or not scalar inferences generally incur a processing cost have yielded conflicting results. We test a prediction made by constraint-based accounts, which specifically seek to unify these conflicting results: that the probability of an interpretation and the speed with which it is processed depends on the contextual support it receives. We manipulated contextual support for the scalar inference in two truth-value judgment experiments by manipulating a lexical feature (presence of partitive "of the") and a pragmatic feature (the implicit Question Under Discussion). Participants' responder type – whether their majority response was pragmatic (reflecting the inference) or literal (reflecting its absence) – was the main predictor of response times: pragmatic responses were faster than literal responses when generated by pragmatic responders; the reverse was true for literal responders. We interpret this as further evidence against costly-implicature accounts and in support of constraint-based accounts of pragmatic processing.

**Keywords:** psycholinguistics; experimental pragmatics; scalar implicature; Question Under Discussion

## Introduction

Listeners routinely go beyond the literal information encoded in the signal to pragmatically infer the speaker's intended meaning. That listeners rapidly draw pragmatic inferences during online processing is well established, but a question that has plagued the literature is whether or not these inferences typically involve a processing cost compared to the processing of literal content (Bott & Noveck, 2004; ?, ?; Huang & Snedeker, 2008; ?, ?; Grodner, Klein, Carbary, & Tanenhaus, 2010; Breheny, Ferguson, & Katsos, 2013; Degen & Tanenhaus, 2016; ?, ?, ?). This question has been prominently addressed for the case of scalar inferences, whereby a listener takes a speaker who produces a sentence like *Jane ate some of the cookies* to mean that she did not eat all of them. The inference proceeds by listeners reasoning that a cooperative speaker should have produced the more informative *Jane ate all of the cookies*, if indeed that alternative sentence is true (according to the speaker) and relevant. The speaker's use of the weaker form, then, implicates the negation of this stronger sentence (?, ?).

The past two decades have seen a wealth of studies from many different experimental paradigms addressing the question of whether or not scalar inferences generally incur a processing cost, with conflicting results. Early studies found that processing sentences that resulted in the inference incurred longer response times (Bott & Noveck, 2004; ?, ?; Degen & Tanenhaus, 2015), longer reading times (?, ?), and led to delays in eye movements to target regions of displays that required the inference be drawn (Huang & Snedeker, 2008; ?, ?; Degen & Tanenhaus, 2016), compared to the processing of literal controls. Later studies found no such delay, especially in eye movement paradigms (Grodner et al., 2010; Breheny et al., 2013; Degen & Tanenhaus, 2016; ?, ?).

Empirically, this conflicting set of results has spurred the development of studies seeking to understand the contextual conditions that facilitate scalar inferences (Zondervan, 2010; Degen, 2015; Augurzky, Franke, & Ulrich, 2019; Marty & Chemla, 2013; Degen & Goodman, 2014). On the theoretical side, it has led to a unification attempt in Degen & Tanenhaus, 2015's constraint-based account. The core tenet of the account is that listeners integrate multiple probabilistic contextual cues to speaker meaning during language processing [jd: mention generative/data explanation approaches?]. Thus, rather than generally incurring a processing cost or generally not incurring a processing cost, the processing effort required to compute an inference is treated as variable. Here, we test the main prediction made by Degen and Tanenhaus (2015)'s constraint-based account: that **the probability of an interpretation and the speed with which it is processed is a function of the contextual support it receives**.

This prediction has previously been tested and borne out in eye movements (Degen & Tanenhaus, 2016), where contextual support for the inference was manipulated via the presence or absence of number terms within the context of the experiment. The inference was processed without a delay relative to literal controls when number terms were absent, but with a delay when the listener had reason to believe that the speaker could have used a more informative number term instead of *some* (Degen & Tanenhaus, 2016), [jd: especially for listeners who generally employed a pragmatic response strategy – mention? mention also the Sun & Breheny follow-up work?].

Here, we extend the investigation of the prediction to a different processing measure – response times within a truth-value judgment task – and a different and more direct way of manipulating the inference's contextual support. We manipulate contextual support via two features: a pragmatic feature – the salient Question Under Discussion (QUD, ?, ?) – and a lexical feature – whether *some* occurs in its partitive form (e.g., *You got some of the gumballs*) or in its non-

partitive form (*You got some gumballs*). Both of these features have previously been shown to modulate scalar inferences from *some* to *not all* (Zondervan, 2010; Degen & Goodman, 2014; Degen, 2015; Degen & Tanenhaus, 2015). The manipulation of these features thus serves two purposes: first, it serves to replicate previous findings showing that these features provide varying contextual support for the scalar inference. Second, establishing varying contextual support allows us to derive predictions about response time patterns under the constraint-based and costly implicature accountns.

- [jd: walk through example within the tvjp to introduce paradigm and multiple interpretations]
- [jd: while the contextual conditions that facilitate inference from 'some' to 'not all' have been increasingly well studied in the past few years – showingn effects of contextual features x, y, and z on si – the extent to which processing speed is a function of contextual conditions is still very much up in the air]
- [jd: we test the constraint-based account by manipulating features of context that provide variable support for the inference and test whether contextual support predicts response times]
- [jd: high-level description of the paradigm and the two contextual features]

In contrast, if scalar inferences generally incur a processing cost, pragmatic responses reflecting that the scalar inference was drawn should be slower to process than literal responses regardless of context. To test the constraint-based versus the costly inference account, we manipulated two features of context between participants in a truth-value judgement task: one lexical (*presence of partititve "of"*) and one pragmatic (*implicit QUD, see (1) and (2)*). This allowed us to obtain estimates of inference rate and processing speed. We further considered a participants *responder type* – whether they have a preference to respond literally or pragmatically – as a predictive feature for response times. While the partitive and the QUD have previously been shown to affect the probability of drawing a scalar inference (Zondervan, 2010; Degen, 2015; Degen & Goodman, 2014; Degen & Tanenhaus, 2015), contextual and participant-specific effects on processing speed have remained under-explored.

This paper is structured as follows: first, we introduce the experimental paradigm, where we employ a truth-value judgment task within the gumball paradigm as introduced by Degen & Tanenhaus, 2015. We then report two experiments conducted within the paradigm, which both manipulated the experiment-wide QUD. The experiments differed in whether the sentences heard on critical trials increased (Exp. 1, partitive *some of*) or decreased (Exp. 2, non-partitive *some*) support for the inference. Both the pragmatic and the lexical feature modulated inference rate, but response time was modulated by the general response strategy participants adopted (which was modulated by the experimental manipulations).[jd: maybe don't foreshadow the results?]

[jd: make connection with rsa / other probabiilistic pragmatics approaches?]

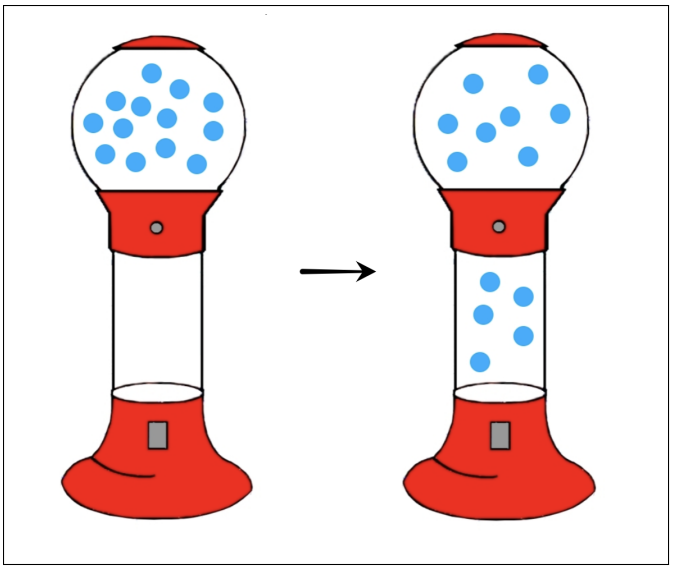

Figure 1: Example display from gumball paradigm. Left: initial display. Right: display with 5 gumballs dropped.

## Experimental paradigm

In both experiments, participants' interpretations were probed using the gumball paradigm introduced by Degen and Tanenhaus (2015). On each trial, participants saw a display of a gumball machine with 13 gumballs in the upper chamber and an empty lower chamber. After 4 seconds, some number of gumballs moved to the lower chamber and a voice reported how many gumballs were distributed (Fig. 1). This pre-recorded statement was of the form "You got X gumballs", where X was a quantifier (*some (of the)*, *all of the*, *none of the*, or a number between 1 and 13). The number of gumballs that dropped to the lower chamber and the quantifier varied (see Table 1).

Participants were assigned to one of the two conditions (*all*-QUD, *any*-QUD) which differed in the cover story they were presented with at the beginning of the experiment (see Table 2). These cover stories were designed to establish the following implicit QUDs:

- ***all*-QUD:** Is the machine empty? → Did I get all of the gumballs? (*more supportive of scalar implicature*)
- ***any*-QUD:** Is the machine jammed? → Did I get any of the gumballs? (*less supportive of scalar implicature*)

## Experiment 1: Partitive statement

In experiment 1 we tested whether the QUD, as a contextual feature of an utterance, could modulate the probability of a scalar implicature and the speed with which it is processed.

| | Set size | | | | | | |
|---|---|---|---|---|---|---|---|
| **Quantifier** | 0 | 2 | 5 | 8 | 11 | 13 | **Total** |
| *some/some of* | 4 | 1 | 1 | 1 | 1 | 8 | 16 |
| *all of* | 2 | 1 | 2 | 1 | 2 | 8 | 16 |
| *none of* | 4 | 1 | 0 | 1 | 1 | 1 | 8 |
| number | 3 | 7 | 7 | 7 | 5 | 3 | 32 |
| **Total** | 13 | 10 | 10 | 10 | 9 | 20 | 72 |

Table 1: Distribution of experimental trials over quantifiers and set sizes.

The main prediction was that in the *all*-QUD condition, the implicit QUD "Did I get all of the gumballs", would be more relevant to the "You got all of the gumballs" interpretation. Thus, there would be more pragmatic "disagree" responses in the critical trials when participants hear "You got some of the gumballs" and get all 13 gumballs. We also predicted that the relevance of the QUD would increase the speed of pragmatic responses and slow down the literal responses.

Procedure, materials, analyses and exclusions were preregistered on OSF and will be available upon publication along with data and experiment scripts.

## Methods

**Participants** We recruited 800 participants on Amazon Mechanical Turk. Participants were required to have a US-based IP address and a minimal approval rating of 95%. They were paid $2.30 (approximately $14/hr).

**Materials and procedure** After reading the cover story of their QUD condition, participants went through a scripted demonstration that showed the consequences of store worker's responses to various scenarios. To ensure that they paid attention to the cover story, they were asked a multiple-choice question about the condition under which the store worker will be fired. When participants answered this question incorrectly, they were presented with the cover story again and had to repeat the demonstration. Halfway through the experiment, participants were asked to answer the multiple-choice question again. This was done to prevent the decay of the implicit QUD over time.

There were 4 practice trials with *all* and *none*. On half of these trials, the statements were correct, and on the other half they were incorrect. After the practice trials, there were 72 experimental trials (see Table 1). On 32 of the trials, the expected answer was yes, and on 32 of the trials, the expected answer was no. The remaining 8 trials were occurances of the critical trial and the main focus of this experiment. On these trials, all 13 gumballs dropped to the lower chamber and participants heard the partitive statement "You got *some of* the gumballs". When participants press YES to agree with this statement, they interpret it semantically as "You got some, and possibly all, of the gumballs" and when they press NO to disagree, they interpret it pramatically as "You got some, but not all, of the gumballs".

**Exclusions** We excluded participants who were self-reported non-native English speakers (n=26), participants who got the second cover story comprehension questions wrong more than twice (n=21) and participants with accuracy lower than 85% on non-critical trials (n=185). Only responses to critical trials are reported below. These exclusions had no effect on the results discussed below. [lk: check again]

[jd: i inserted the following; can also be excluded] **Analysis and predictions** Only responses on critical trials are analyzed below. We conducted two types of analyses to address the two questions of interest:

1. Does the QUD modulate the probability of drawing a scalar inference?

2. Does the contextual support that an interpretation (either pragmatic or literal) receives

To this end, we conducted a mixed effect logistic regression predicting the log odds of pragmatic "no" over literal "yes" responses.

## Results and discussion

### Judgments

Proportion of pragmatic responses on critical trials are shown in Figure 2. 78% of responses given by the participants in the *all*-QUD condition were pragmatic compared to 71% pragmatic responses given by participants in the *any*-QUD condition. We ran a mixed effects logistic regression predicting response type with the maximal random effects structure justified by the design – random by-participant intercepts – from a fixed effect of QUD. We observed a main effect of QUD such that there were more pragmatic responses in the *all*-QUD condition compared to the *any*-QUD condition ($\beta$=1.31, SE=0.52, $p<.05$).

### Analysis of Variability in Judgments

Figure 3 shows the distribution of participants over number of pragmatic responses given on critical trials. Participants who either gave 0 or 8 pragmatic responses were completely consistent in their responses (62%, of which 20% completely literal and 80% completely pragmatic). Figure 3 mirrors Figure 3 [lk: can make the plots seperately] and shows that the distribution of pragmatic responses in the *all*-QUD condition is shifted towards the more pragmatic end of the continuum compared to the *any*-QUD condition. Thus, while some participants were entirely consistent, there was also substantial inter-participant variability in consistency. For the purpose of the subsequent response time analyses, and following previous researchers (Bott & Noveck, 2004; Degen, 2015), we divided participants into two groups: participants with more than 4 pragmatic responses were categorized as *pragmatic* responders (74%) and participants with fewer than 4 pragmatic responses were categorized as *literal* responders (22%). 15 participants (3%) gave an equal number of pragmatic and literal responses and were excluded from the response time

| all-QUD | any-QUD |
|---|---|
| You are at a candy store and are testing a row of gumball machines. These are special gumball machines that say how many gumballs you got. However, this report is sometimes faulty. | |
| The store worker tells you that his boss has threatened to fire him if the gumball machines are left empty, and he really needs this job. He cannot see the machines from the register, but he can normally tell how full they are by the machines' statements. | The store worker tells you that machines sometimes jam and don't deliver any gumballs. His boss has threatened to fire him if the gumball machines stay jammed, and he really needs this job. He cannot see the machines from the register, but he can normally tell if they are working by the machines' statements. |
| He asks you to tell him if the statement is right or wrong, so that he will know if a machine is empty and needs to be refilled. After you hear the statement, you have 4 seconds to notify the store worker, so please make a decision as quickly as possible. | |

Table 2: Cover stories for each QUD condition.

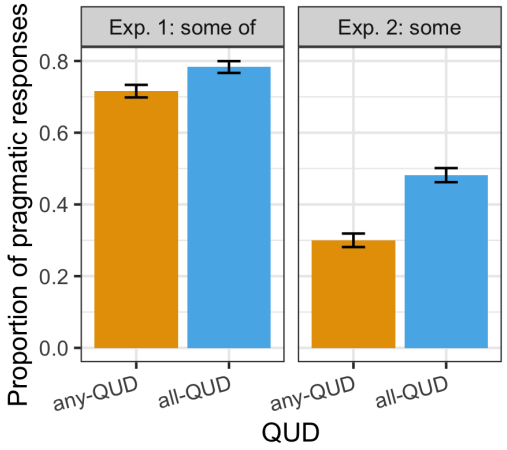


Figure 2: Proportion of pragmatic responses on partitive "some of" (left) and non-partitive "some" (right) critical trials. Error bars indicate bootstrapped 95% confidence intervals.

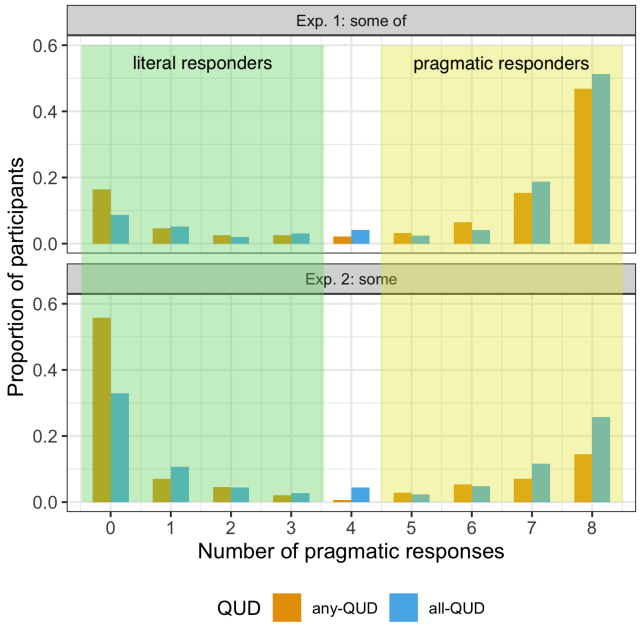


Figure 3: Distribution of participants over number of pragmatic responses given on critical trials. Participants with $< 4$ pragmatic responses were categorized as literal responders (green), participants with $> 4$ responses as pragmatic responders (yellow).

analysis. [jd: perhaps include a note saying what happens if the inconsistent people are included in a reasonable way?]

**Response Times**

We predicted that the speed with which a scalar implicature is processed would increase with more contextual support. To test this hypothesis, we ran a mixed effects linear regression model with random by-participant intercepts predicting log-transformed response time from fixed effects of QUD, response type and their interaction. We found an interaction between QUD and response ($\beta$=-1.12, SE=2.51, t=-4.45, p$<$.0001) such that pragmatic responses were faster under the *all*-QUD than under the *any*-QUD. This shows that the relevance of an implicature to a contextual QUD modulates the speed of implicature processing.

When we added responder type as predictor to this model, the largest observed effect was the interaction between responder type and response ($\beta$=-2.72, SE=3.34, t=-8.14, p$<$.0001), such that pragmatic responses were faster than literal responses for *pragmatic* responders and literal responses were faster than pragmatic responses for *literal* responders (see Figure 4).

## Experiment 2: Non-partitive statement

In Experiment 2 we tested whether the absence of the partitive, previously shown to decrease the contextual support for the inference, would decrease the scalar inference rate. We also investigated whether the QUD and responder type ef-

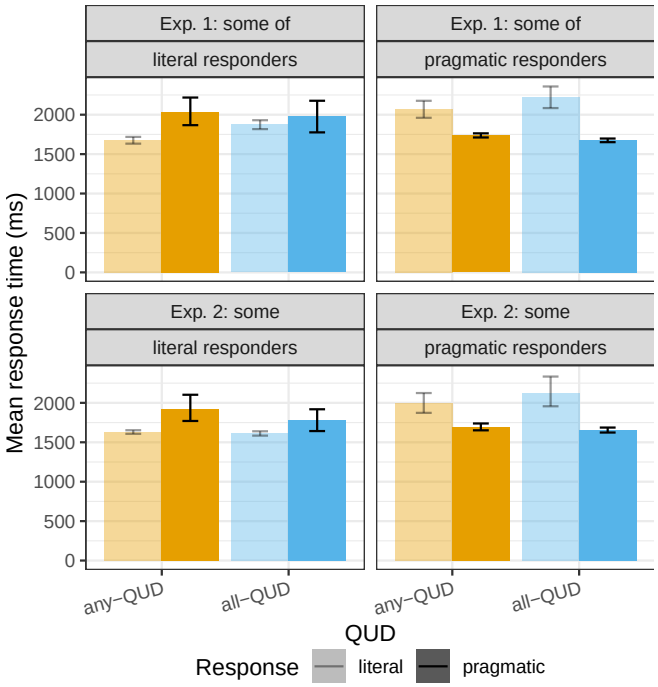


Figure 4: Mean response times for literal (light) and pragmatic (dark) responses generated by literal (left column) and pragmatic (right column) responders on partitive "some of" (top panels) and non-partitive "some" (bottom panels) critical trials.

fects observed in Experiment 1 would replicate without the contextual support of the lexical cue.

## Methods

**Participants** We recruited 800 participants on Amazon Mechanical Turk. Participants had to have a US-based IP address and a minimal approval rating of 95%, and they were paid $2.3 (approximately $14/hr).

**Materials and procedure** The materials and procedures were the same as in Experiment 1 except on critical trials, when all 13 gumballs dropped to the lower chamber, participants heard the non-partitive statement "You got *some* gumballs".

**Exclusions** As in Experiment 1, we excluded non-native English speakers (n=21), participants who got the second comprehension question wrong more than twice (n=15), and participants that had accuracy lower than 85% on non-critical trials (n=189).

**Analysis and predictions** [jd: XXX]

## Results and discussion

### Judgments

We found that overall, participants in Experiment 2 who heard the non-partitive statement were less likely to respond pragmatically compared to participants in Experiment 1 who heard its partitive counterpart (β=7.16, SE=0.69, p<.0001)(see Fig 2), replicating previous studies (Degen & Tanenhaus, 2015; Degen, 2015).

We also found an interaction of the lexical cue and QUD (β=-3.06, SE=0.90, p<.0001) but it was mainly because of the QUD effect being bigger for the non-partitive condition due to a ceiling effect. [lk: fix]

We replicated the QUD effect found in Experiment 1. Participants in the *all*-QUD condition gave more pragmatic "disagree" responses than participants in the *any*-QUD (β=4.69, SE=0.80, p<.0001). [lk: maybe say why we pool data] When we pooled the data from Experiment 1 and 2, QUD remained to be a predictor of response type (β=2.85, SE=0.44, p<.0001).

### Analysis of Variability in Judgments

44% of participants were completly consistent in their pragmatic responses compared to 20% of participants that gave 0 pragmatic responses. 15 participants (3%) were excluded from the response time analysis because they gave equal number of pragmatic and literal responses.

As shown in Fig 3, in the *any*-QUD condition, the distribution of pragmatic responses is shifted towards the more literal end of the continuum compared to the *all*-QUD condition. Overall, we see a shift [lk: complete opposites]

### Response Times

In order to compare the response times of participants from Experiment 1 and 2, we had to calculate each response time with respect to the length of the audio stimuli participants heard. For each experiment, we subtracted the length of the onset of the word "gumball" (Experiment 1: 868ms, Experiment 2: 736ms) from all response times. This ensured that response times [lk: XXX]. We were interested in whether the lack of the partitive would slow down pragmatic responses and speed up literal responses compared to its partitive counterpart. We analyzed the data using a mixed effects linear regression model with by-participant intercepts to predict log-transformed response times. The model included centered fixed effects of quantifier and response. We found a significant interaction between quantifier and response such that (β=-0.10, SE=0.04, t=2.48 p<.05) [lk: explain]

[lk: We weren't able to replicate the QUD effect on the speed of scalar implicatures when the non-partitive form was used.]

[lk: which one do we want to focus on?] Finally, we ran a full model on the pooled data with QUD, quantifier, response and responder type as centered predictors. We found a main effect of quantifier (β=5.24, SE=1.52, t=3.45, p<.001), response (β=-8.69, SE=1.15, t=-7.58, p<.001), an interaction between qud and response (β=-1.06, SE=2.29, t=-4.62, p<.001), and an interaction between response and responder type (β=-2.69, SE=2.30 t=-11.68, p<.001)

## General discussion and conclusion

Contextual factors affect listeners' overall contextual response strategy which in turn impacts the speed with which they process the preferred interpretation. This is evidence

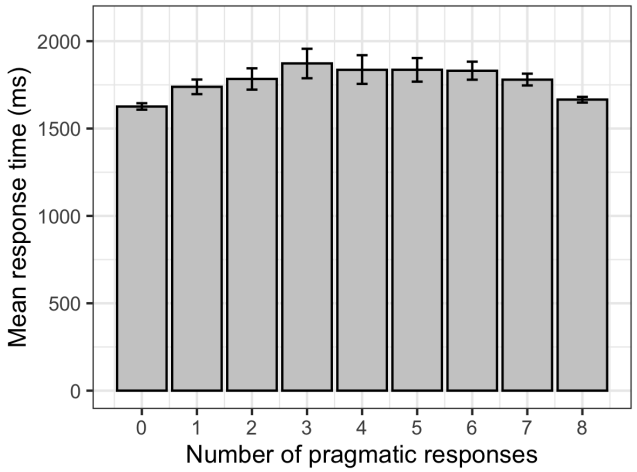


Figure 5: Mean response times for participants grouped based on the number of pragmatic responses they gave.

against costly inferece accounts and in support of constraint-based accounts.

## References


Augurzky, P., Franke, M., & Ulrich, R. (2019, 08). Gricean expectations in online sentence comprehension: An erp study on the processing of scalar inferences. *Cognitive Science*, *43*. doi: 10.1111/cogs.12776

Bott, L., & Noveck, I. (2004, 10). 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

Breheny, R., Ferguson, H. J., & Katsos, N. (2013). Investigating the timecourse of accessing conversational implicatures during incremental sentence interpretation. *Language and Cognitive Processes*, *28*(4), 443-467. doi: 10.1080/01690965.2011.649040

Degen, J. (2015, May). Investigating the distribution of *some* (but not *all*) implicatures using corpora and web-based methods. *Semantics and Pragmatics*, *8*(11), 1–55. doi: 10.3765/sp.8.11

Degen, J., & Goodman, N. (2014). Lost your marbles? the puzzle of dependent measures in experimental pragmatics. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

Degen, J., & Tanenhaus, M. K. (2015). Processing scalar implicature: A constraint-based approach. *Cognitive science*, *39*(4), 667–710.

Degen, J., & Tanenhaus, M. K. (2016). Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive science*, *40*(1), 172–201.

Grodner, D., Klein, N., Carbary, K., & Tanenhaus, M. (2010, 07). "some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition*, *116*, 42-55. doi: 10.1016/j.cognition.2010.03.014

Huang, Y.-t., & Snedeker, J. (2008, 11). Online interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cognitive psychology*, *58*, 376-415. doi: 10.1016/j.cogpsych.2008.09.001

Marty, P., & Chemla, E. (2013, 07). Scalar implicatures: Working memory and a comparison with only. *Frontiers in psychology*, *4*, 403. doi: 10.3389/fpsyg.2013.00403

Zondervan, A. (2010, 01). Scalar implicatures or focus: an experimental approach.