Probability and processing speed of scalar inferences is context-dependent

Leyla Kursat and **Judith Degen**
{lkursat, jdegen}@stanford.edu
Department of Linguistics, Stanford University
Stanford, CA 94305, USA

Abstract

Studies addressing the question of whether scalar inferences generally incur a processing cost have yielded conflicting results. Constraint-based accounts, which seek to unify these conflicting results, make a prediction which we test here: 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 inference accounts and in support of constraint-based accounts of pragmatic processing.

Keywords: psycholinguistics; experimental pragmatics; scalar inference; 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; Breheny, Katsos, & Williams, 2006; Huang & Snedeker, 2009, 2011; Grodner, Klein, Carbary, & Tanenhaus, 2010; Breheny, Ferguson, & Katsos, 2013; Degen & Tanenhaus, 2016; De Neys & Schaeken, 2007; Tomlinson Jr, Bailey, & Bott, 2013). 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 standard account of the inference is that listeners reason that a cooperative speaker should have produced the more informative *Jane ate all of the cookies*, if indeed that alternative sentence was true (according to the speaker) and relevant. The speaker's use of the weaker form, then, implicates the negation of this stronger sentence (Grice, 1975).

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 evidence consistent with a *costly inference account*. Under such an account, computing the inference is assumed to be a cognitively effortful process because it requires processing pragmatic information in addition to the literal semantics of the sentence (Huang & Snedeker, 2009; Tomlinson Jr et al., 2013). This additional processing is taken to be effortful. Indeed, this account is supported by studies in which processing sentences that resulted in the inference incurred longer response times (Bott & Noveck, 2004; Tomlinson Jr et al., 2013; Degen & Tanenhaus, 2015), longer reading times (Breheny et al., 2006), and delays in eye movements to target regions of displays that required the inference be drawn (Huang & Snedeker, 2009, 2011; Degen & Tanenhaus, 2016), compared to processing sentences literally. However, other studies have found no such delay, especially in eye movement paradigms (Grodner et al., 2010; Breheny et al., 2013; Degen & Tanenhaus, 2016; Sun & Breheny, 2019).

Empirically, this conflicting set of results has spurred the development of studies seeking to understand the contextual conditions that facilitate scalar inferences (Zondervan, 2010; Bonnefon, Feeney, & Villejoubert, 2009; Degen, 2015; Augurzky, Franke, & Ulrich, 2019; Marty & Chemla, 2013; Degen & Goodman, 2014; Sun & Breheny, 2019). On the theoretical side, the results suggest the need for a unified theory of the conditions under which processing delays (fail to) arise. The *constraint-based account* proposed by Degen & Tanenhaus, 2015 is such an account. The core tenet of the account is that listeners integrate multiple probabilistic contextual cues to speaker meaning during language processing and that it is not the integration of pragmatic information per se that is costly, but rather the processing of the inference in contexts where support for it is weak. Thus, rather than generally incurring a processing cost or generally not incurring a processing cost, the processing effort required to compute an inference may vary. Here, we test the main prediction made by the 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).

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, Roberts, 2012) – 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 inference accounts.

We proceed as follows: first, we introduce the experimental paradigm, a truth-value judgment task within the gumball paradigm as introduced by Degen and Tanenhaus (2015) and lay out the predictions of the competing theoretical accounts in detail. We then report two experiments conducted within the paradigm, which both manipulated the experiment-wide QUD. The experiments differed in whether sentences heard on critical trials increased support for the inference (Exp. 1, partitive *some of*) or decreased it (Exp. 2, non-partitive *some*). Both the pragmatic and the lexical feature modulated inference rate, replicating previous results. Our novel contribution lies in the response time analyses, which we discuss with respect to theoretical accounts of interest below.

Experimental paradigm and predictions

In both experiments, participants' interpretations were probed using the gumball paradigm introduced by Degen and Tanenhaus (2015), which builds on earlier truth-value judgment work (Bott & Noveck, 2004). On critical trials, participants heard a sentence like *You got some of the gumballs* as a description of a set of facts that made the stronger alternative *You got all of the gumballs* true and were asked whether or not they agree with the statement. If participants interpreted the utterance literally (*You got at least some of the gumballs, and possibly all of them*), they responded "agree". If, instead, they interpreted the utterance pragmatically (*You got some, but not all, of the gumballs*), they responded "disagree".

Under the costly inference account, literal responses should always be faster than pragmatic responses, regardless of the contextual information participants are provided with. This is indeed the result reported by Bott & Noveck, 2004. In contrast, under the constraint-based account, the more the context supports the inference (as measured in proportions of pragmatic responses), the faster participants should be to provide a pragmatic response and the slower they should be to provide a literal response. Conversely, the weaker the contextual support for the inference, the slower the pragmatic response should be, and the faster the literal one. The strongest test of the constraint-based account would be evidenced in a response time pattern whereby pragmatic responses are faster than literal responses, which to our knowledge has not been previously demonstrated. Note an attractive feature of the constraint-based account: it links the processing effort involved in computing both pragmatic *and* literal interpretations to the contextual support for the interpretation, rather than treating the processing of literal content as a monolith.

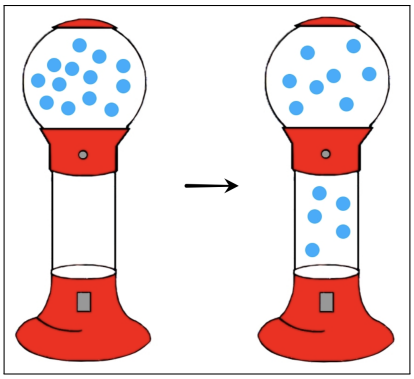

Figure 1: Example display from gumball paradigm. Left: initial display. Right: display with 5 gumballs dropped.

In studies that use truth-value judgment tasks to probe scalar inferences, participants typically complete multiple critical trials of the kind described above (Noveck & Posada, 2003; Bott & Noveck, 2004; Degen & Tanenhaus, 2015; Jasbi, Waldon, & Degen, 2019). Participants don't always respond consistently across critical trials (see Degen & Tanenhaus, 2015, for a detailed response consistency analysis), which has led researchers to categorize participants as either "pragmatic" or "literal" responders and conduct responder type analyses to investigate whether participants' varying response strategies result in varying processing strategies. For instance, Degen & Tanenhaus, 2016 found that pragmatic responders were more strongly affected by the experimental manipulation of number term presence than literal responders, who tended to wait for disambiguating information before fixating on target regions of the display. We include such responder type analyses in our response time analyses. Specifically, we treat responder type as an additional variable that may influence processing times by providing latent contextual support for one over another interpretation: if being a pragmatic responder increases the latent contextual support for the pragmatic response, these responses should be faster than literal responses; conversely for literal responders.

Within each experiment, we followed Degen (2013) in manipulating contextual support for the inference via an experiment-level implicit QUD invoked by a cover story that either made the stronger alternative more relevant (*all*-QUD

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.	He asks you to tell him if the statement is right or wrong, so that he will know if a machine isn't working and need to be fixed.
After you hear the statement, you have 4 seconds to notify the store worker, so please make a decision as quickly as possible.	

Table 1: Cover stories for each QUD condition.

condition, more support for scalar inference) or less relevant (*any*-QUD condition, less support for scalar inference).[1]

Experiment 1: Partitive sentences

In Exp. 1 we tested whether the QUD modulates the probability of a scalar inference and whether the QUD and participants' responder type jointly modulate the speed with which pragmatic and literal interpretations are processed. Sentences on critical trials included *some* in its partitive form, previously shown to support the inference (Degen, 2015).

Methods

Participants, materials, procedure. We recruited 800 participants on Amazon's Mechanical Turk. 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 (Fig. 1) and a voice reported how many gumballs were distributed. Participants' task was to indicate whether they 'agree' or 'disagree' with the statement by pressing the F or J key as quickly as possible. The 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 quantifier and the number of gumballs that dropped to the lower chamber varied.

Before proceeding to the main body of the experiment, participants read a cover story to induce an implicit QUD (see Table 1 for cover stories). They completed a scripted demonstration that introduced a gumball store worker who will be fired if he either fails to re-fill a gumball machine when it's empty (*all*-QUD condition) or if he fails to fix a machine if it's not dispensing gumballs (*any*-QUD condition). To ensure that participants paid attention to the cover story, they were asked a multiple-choice question about the condition under which the store worker will be fired. If participants answered this question incorrectly, they were presented with the cover story again and repeated 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.

	Set size						
Quantifier	0	2	5	8	11	13	**Total**
some of/some	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 2: Distribution of trials over quantifiers and set sizes.

There were 4 practice trials with *all* and *none*. On half of these trials the statements were correct, on the other half they were incorrect. After practice trials, participants completed 72 experimental trials (see Table 2). On 32 trials, the expected answer was 'agree', on another 32 trials, the expected answer was 'disagree'. The remaining 8 trials were critical trials on which all 13 gumballs dropped to the lower chamber and participants heard the partitive scalar statement *You got some of the gumballs*. To reiterate, if participants pressed J to agree with the statement, they interpreted it literally; if they pressed F to disagree, they interpreted it pragmatically.

Exclusions. We excluded participants who were self-reported non-native English speakers (n=20), participants who incorrectly answered the second comprehension question more than twice (n=17) and participants with accuracy lower than 85% on non-critical trials (n=235). Only responses to critical trials are reported below. These exclusions had no qualitative effect on the results discussed below.

Analysis. Only responses on critical trials were analyzed here and in Exp. 2. We conducted three types of analyses to address the questions of interest. First, we analyzed judgments to test whether the contextual QUD modulated scalar

[1]Procedure, materials, analyses and exclusions were pre-registered at https://osf.io/xkh8g (Exp. 1) and https://osf.io/49uqm (Exp. 2). The collected sample size for Exp. 1 (800 participants) was much larger than the originally pre-registered sample size (100 participants) because a preliminary power analysis suggested that the pre-registered sample size would not yield adequate power.

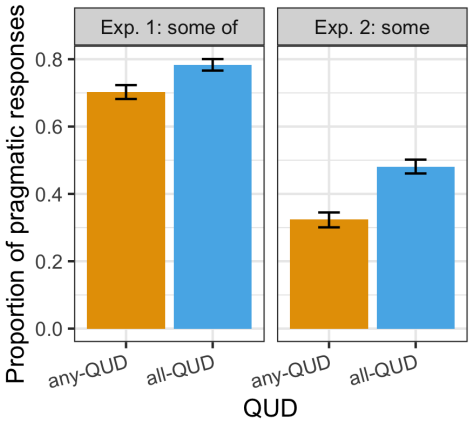

Figure 2: Proportion of pragmatic responses on partitive *some of* (Exp. 1, left) and non-partitive *some* (right, Exp. 2) critical trials. Error bars indicate bootstrapped 95% confidence intervals.

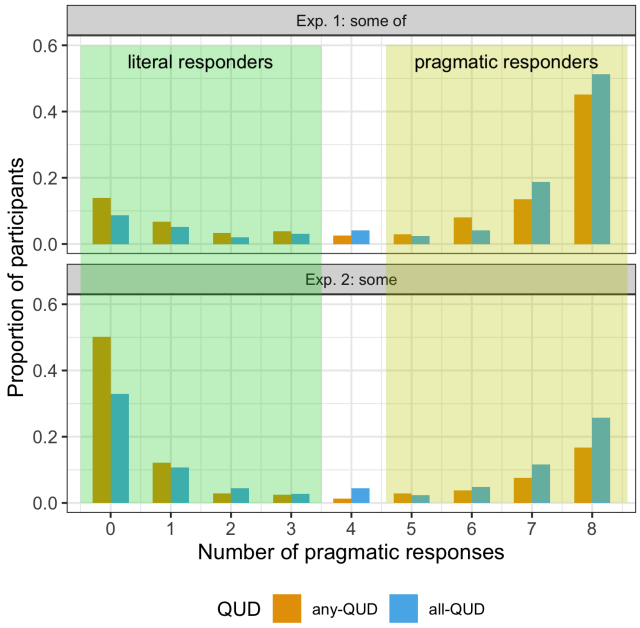

Figure 3: Distribution of participants over number of pragmatic responses given on critical trials. Participants with < 4 pragmatic responses were categorized as literal responders (green), participants with > 4 responses as pragmatic responders (yellow).

inferences. To this end, we conducted mixed effects logistic regression predicting the log odds of a pragmatic over a literal response from a fixed effect of QUD and the maximal random effects structure justified by the design – random by-participant intercepts. Second, we analyzed participants' response consistency and categorized them as literal or pragmatic responders. Third, we tested the prediction made by the constraint-based account that responses that reflect a particular interpretation (pragmatic or literal) should be processed more quickly, the greater the contextual support for the interpretation. To this end, we conducted a mixed effects linear regression model predicting log-transformed response time from fixed effects of QUD, response type, and responder type, with the maximal random effects structure justified by the design – random by-participant intercepts and slopes for response type. All fixed effect predictors were centered before entering their respective analysis.

Results and discussion

Judgments. Proportion of pragmatic responses on critical trials are shown on the left in Figure 2. We observed a main effect of QUD such that there were more pragmatic responses in the *all*-QUD condition (78%) compared to the *any*-QUD condition (70%, β=1.27, SE=0.53, p<.05), replicating previous QUD effects on scalar inference (Degen & Goodman, 2014; Zondervan, 2010).

Analysis of variability in judgments. The top panel of Fig. 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 (60%, of which 19% completely literal and 81% completely pragmatic). Fig. 3 reflects Fig. 2 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 & Tanenhaus, 2015), we divided participants into two groups: participants with more than 4 pragmatic responses were categorized as *pragmatic* responders (73%) and participants with fewer than 4 pragmatic responses were categorized as *literal* responders (23%). 18 participants (3%) gave an equal number of pragmatic and literal responses and were excluded from the response time analysis.

Response times. Response times are shown in the top panels of Fig. 4. We observed an interaction between QUD and response (β=-8.16, *SE*=3.40, *t*=-2.40, *p*<.05). Simple effects analysis revealed that this interaction was driven by literal responses being slower under the *all*-QUD than under the *any*-QUD (β=0.09, SE=0.03, t=2.58, p<.05), while there was no QUD-based difference in response times for pragmatic responses (β=0.01, SE=0.05, t=0.11, p<.91). We also observed an interaction between responder type and response (β=-2.37, SE=3.25, t=-7.29, p<.0001). Simple effects analysis revealed that this interaction was driven by pragmatic responses being faster than literal responses for *pragmatic* responders (β=-0.17, SE=0.03, t=-5.53, p<.0001) and pragmatic responses being slower than literal responses for *literal* responders (β=0.07, SE=0.03, t=1.98, p<.05).

While we did not observe a direct effect of the QUD on

pragmatic response time, these results nevertheless provide evidence against the costly inference account: rather than pragmatic responses being generally slower than literal responses, they were only so when generated by literal responders. In addition, the QUD modulated literal response time in the direction predicted by the constraint-based account – when support for the literal interpretation was weaker, response time increased.

These results also excitingly show that pragmatic responses can be faster than literal responses under certain conditions, namely when produced by pragmatic responders. These results are consistent with the constraint-based account. To test whether the contextual effects remain stable when overall decreasing the contextual support for the inference, we conducted Exp. 2.

Experiment 2: Non-partitive sentences

Exp. 2 was identical to Exp. 1, but the sentence on critical trials was the non-partitive *You got some gumballs*, previously shown to yield lower inference rates than the partitive version.

Methods

Participants, materials, procedure. We recruited 800 participants on MTurk. The materials and procedure were identical to Exp. 1 except on critical trials, where participants heard the non-partitive statement *You got some gumballs*.

Exclusions. We excluded non-native English speakers (n=19), participants who got the second comprehension question wrong more than twice (n=29), and participants that had accuracy lower than 85% on non-critical trials (n=217).

Results

Judgments. Proportion of pragmatic responses on critical trials are shown on the right in Figure 2. We replicated the QUD effect found in Exp. 1: participants in the *all*-QUD condition gave more pragmatic "disagree" responses (48%) than participants in the *any*-QUD condition (30%, β=3.07, SE=0.63, p<.0001). However, as is evident from Fig. 3, the rate of pragmatic responses was greatly reduced overall compared to Exp. 1. This was borne out in a mixed effects logistic regression where the data from both experiments was pooled and sentence form added as a fixed effect predictor (β=5.89, SE=0.55, p<.0001), replicating previous studies (Degen & Tanenhaus, 2015; Degen, 2015).

Analysis of variability in judgments. The bottom panel of Fig. 3 shows the distribution of participants over number of pragmatic responses given on critical trials. Reflecting the average changes in Fig 3, overall and in the *any*-QUD condition, the distribution of responses shifted towards the more literal end of the continuum compared to Exp. 1 and the *all*-QUD condition. 22% of participants were completely consistent in providing pragmatic responses compared to 40% of participants who consistently responded literally. Overall, 38% of participants were categorized as pragmatic responders and 58% as literal responders. 16 participants (3%) were excluded from the response time analysis because they gave equal number of pragmatic and literal responses.

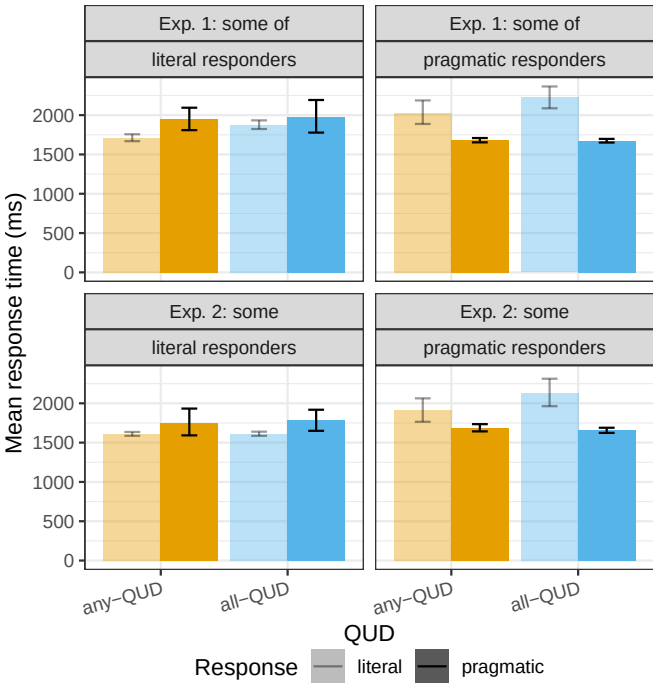

Figure 4: Mean response times for literal (light) and pragmatic (dark) responses generated by literal (left) and pragmatic (right) responders on partitive *some of* (Exp. 1, top) and non-partitive *some* (Exp. 2, bottom) critical trials.

Response times. Response times are shown in the bottom panels of Fig. 4. We observed an interaction between responder type and response (β=-0.27, SE=0.03, t=-8.38, p<.0001), which simple effects analysis revealed was driven by pragmatic responses being faster than literal responses for *pragmatic* responders (β=-0.16, SE=0.03, t=-4.91, p<.0001) and pragmatic responses being slower than literal responses for *literal* responders (β=0.09, SE=0.03, t=2.75, p<.007), replicating the result of Exp. 1.

However, in contrast to Exp. 1, instead of a two-way interaction between QUD and response, we observed a three-way interaction between QUD, response, and responder type (β=-0.19, *SE*=0.07, *t*=-2.84, *p*<.01). Simple effects analysis revealed this interaction was driven by the two-way interaction between QUD and response only being significant for pragmatic responders (β=-0.15, *SE*=0.05, *t*=-2.95, *p*<.01), but not for literal responders (β=0.04, *SE*=0.04, *t*=0.93, *p*<.36). The two-way interaction for pragmatic responders was driven by literal responses being slower in the *all*-QUD condition than in the *any*-QUD condition (β=0.12, *SE*=0.06, *t*=2.16, *p*<.05), while there was no evidence that QUD modulated pragmatic response time (β=-0.03, *SE*=0.03, *t*=-0.96, *p*<.34).

Response time comparison between Exps. 1 and 2. Across both experiments we replicated the result that pragmatic responses were faster than literal responses when generated by

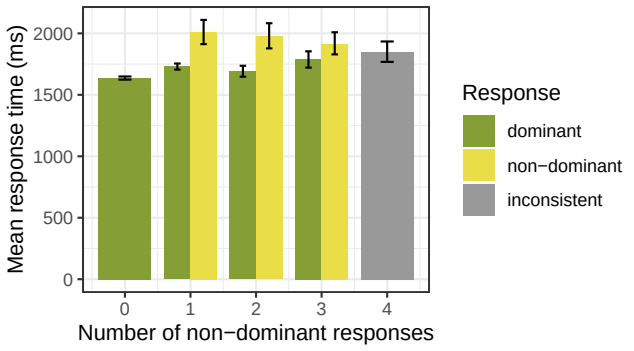

Figure 5: Mean response time as a function of response dominance and by-participant response consistency: 0 non-dominant responses indicates consistently literal/pragmatic responders; 4 non-dominant responses indicates entirely inconsistent responders.

pragmatic responders and vice versa for literal responders. We also found that literal response speed was generally modulated by the QUD such that literal responses were slower where the literal interpretation received less support, as established by the judgment data. There was no evidence that the QUD directly modulated pragmatic response speed.

It would have been interesting to assess the effect of the partitive form on response times across the two experiments – if the strong version of the constraint-based account is right, the partitive should lead to faster pragmatic responses than the non-partitive form. Unfortunately, directly comparing the response time results from Exps. 1 and 2 is not possible because of the different overall lengths of the audio files, which is reflected in the response times (partitive utterances were on average 132ms longer than non-partitive ones).

General discussion and conclusion

The issue of whether or not scalar inferences generally incur a processing cost has been the subject of much debate. Here we have added an additional data point against costly inference accounts: we identified conditions under which pragmatic responses are provided more quickly than literal responses, to our knowledge the first time the reversal of this classic response time pattern has been shown. We have also replicated previous findings showing that both a QUD that makes the stronger alternative more contextually relevant and the presence of the partitive increase the rate of scalar inferences.

Under the strong constraint-based prediction, response times associated with the literal and pragmatic interpretation should have been directly related to the amount of contextual support they received. Instead, rather than a general modulation of response times by the QUD, we observed the predicted difference only for literal responses. However, the strongest effect on response times was the interaction between responder type and response. That pragmatic responders provided faster pragmatic than literal responses and vice versa for literal responders suggests that, rather than directly affecting response times, contextual cues to pragmatic meaning may guide listeners' overall expectations for likely meanings, which in turn affect how much processing effort must be invested to arrive at a particular response. We thus interpret these results as further limited evidence for constraint-based accounts of pragmatic processing.

It is possible that the relatively coarse-grained response time measure obscured an existing underlying inference cost, and that response times are simply a function of multiple processes, only one of which is the computation of the inference (see Feeney, Scrafton, Duckworth, & Handley, 2004; Huang & Snedeker, 2009, for arguments to this effect). For instance, response times might also reflect the cost of verifying that the computed meaning is contextually satisfied.[2] Thus, a weak version of the costly inference account that allows for response times to reflect more than just inference computation cost is not ruled out by the reported results.

This points to a general methodological issue with the use of truth-value judgment tasks in experimental pragmatics: to decide between accounts of implicature calculation based on response times in such experiments, explicit models that link response times and probabilities to theories of pragmatic inference are necessary (see also Jasbi et al., 2019). While there is a general dearth of such explicit linking functions, recent work in probabilistic pragmatics provides an alternative linking hypothesis from inference probability to truth-value judgment probability: Waldon and Degen (2020) propose to treat behavioral responses as the result of probabilistic reasoning about speakers' likely productions. Extending this account to response times is a promising way forward.

The complex causal links between contextual cues to meaning and inferential processing effort require further investigation, but it is clear from this and previous work that treating scalar inferences as monoliths with a particular associated inferential effort is unlikely to yield a satisfying theory of pragmatic processing.

[2]Indeed, a reviewer asks whether the interaction between response and responder type might be due to different verification strategies employed by the two groups, whereby pragmatic responders 'precode' the pragmatic interpretation and focus only on whether or not gumballs are left in the top part of the machine, and literal responders 'precode' the literal interpretation and focus only on whether there are at least some gumballs in the lower part of the machine, thus making the dominant response – whether literal or pragmatic – faster than the non-dominant response. This predicts that participants' dominant responses should be uniformly fast, and non-dominant responses uniformly slow. This was not the case: while dominant response were generally faster than non-dominant responses, there was a gradient effect of the amount of by-participant response consistency on response time (see Fig. 5): response time increased with increasing response inconsistency for the dominant response (β=0.02, SE=0.01, t=3.61, $p<.001$) and decreased with increasing response inconsistency for the non-dominant response (β=-0.03, SE=0.006, t=-2.37, $p<.05$), independent of response type (literal or pragmatic). Default verification strategies cannot explain this pattern. A way of interpreting these results, following Degen and Tanenhaus (2015), is that inconsistency in results reflects uncertainty about the QUD, and that the observed slowdown with increasing response inconsistency reflects a cost associated with maintaining uncertainty about the contextual QUD.

References

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

Bonnefon, J.-F., Feeney, A., & Villejoubert, G. (2009). When some is actually all: Scalar inferences in face-threatening contexts. *Cognition*, *112*(2), 249–258.

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

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

Breheny, R., Katsos, N., & Williams, J. (2006). Are generalised scalar implicatures generated by default? an on-line investigation into the role of context in generating pragmatic inferences. *Cognition*, *100*(3), 434–463.

Degen, J. (2013). *Alternatives in pragmatic reasoning*. University of Rochester.

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

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.

De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental psychology*, *54*(2), 128–133.

Feeney, A., Scrafton, S., Duckworth, A., & Handley, S. J. (2004). The story of some: Everyday pragmatic inference by children and adults. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *58*(2), 121.

Grice, H. P. (1975). Logic and conversation. *Syntax and Semantics*, 41–58.

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

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

Huang, Y. T., & Snedeker, J. (2011). Cascading activation across levels of representation in children's lexical processing. *Journal of Child Language*, *38*(3), 644–661.

Jasbi, M., Waldon, B., & Degen, J. (2019). Linking hypothesis and number of response options modulate inferred scalar implicature rate. *Frontiers in psychology*, *10*.

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

Noveck, I. A., & Posada, A. (2003). Characterizing the time course of an implicature: An evoked potentials study. *Brain and language*, *85*(2), 203–210.

Roberts, C. (2012). Information structure: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, *5*, 6–1.

Sun, C., & Breheny, R. (2019). Another look at the online processing of scalar inferences: an investigation of conflicting findings from visual-world eye-tracking studies. *Language, Cognition and Neuroscience*, 1–31.

Tomlinson Jr, J. M., Bailey, T. M., & Bott, L. (2013). Possibly all of that and then some: Scalar implicatures are understood in two steps. *Journal of memory and language*, *69*(1), 18–35.

Waldon, B., & Degen, J. (2020). Modeling behavior in truth value judgment task experiments. *Proceedings of the Society for Computation in Linguistics*, *3*(1), 10–19.

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