Pragmatic inferences are QUD-sensitive: an experimental study[1]

ESZTER RONAI

The University of Chicago

MING XIANG

The University of Chicago

Implicatures serve as an important testing ground for examining the process of integrating semantic and pragmatic information. Starting with Bott & Noveck (2004), several studies have found that implicature computation is costly. More recently, attention has shifted towards identifying contextual cues that modulate this processing cost. Specifically, it has been hypothesised that calculation rate and processing cost are a function of whether the Question Under Discussion (QUD) supports generating the implicature (Degen & Tanenhaus 2015, Degen 2013). In this paper, we present a novel elicitation task establishing what the relevant QUDs are for a given context (Experiment 1). In Experiment 2, a sentence-picture verification study, we extend earlier findings about the effect of QUDs on scalar inference to a different kind of pragmatic inference: *it*-cleft exhaustivity. For both inferences, we find that under QUDs that bias towards calculation, there is no increase in reaction times, but under QUDs that bias against calculating the inference we observe longer reaction times. These results are most compatible with a constraint-based account of implicature, where QUD is one of many cues. Additionally, we explore whether our findings can be informative in narrowing down precisely what aspect of the inferential process incurs a cost.

Keywords: experimental pragmatics, Question Under Discussion, implicatures, inference processing

1. Introduction

In natural language communication, hearers regularly infer messages beyond what is literally, explicitly said by the speaker. In doing so, they rely not only on what is said, but also on what is not said. One well-studied instantiation of this phenomenon is (the family of) implicatures, exemplified in (1) by scalar inference.

(1) Mary ate some of the cookies.
Literal: Mary ate some and possibly all of the cookies.
Inference-enriched: Mary ate some but **not all** of the cookies.

[1] For their insightful comments and suggestions, we thank the three anonymous reviewers, Hannah Rohde and John Tomlinson, the audiences at AMLaP 24 and NELS 49, and especially Chris Kennedy and Michael Tabatowski. We are grateful to Bob van Tiel for sharing experimental materials with us and to Zsolt Veraszto for technical help. All mistakes and shortcomings are our own.

To derive the inference-enriched reading of the sentence in (1), hearers consider and reason about the informationally stronger alternative that could have been uttered in place of what has actually been uttered. In particular, hearers are taken to reason that the stronger statement (*Mary ate all of the cookies*) was also available to the speaker, but because she chose not to utter it, its negation can be inferred. This process can be viewed as an interaction of the Quality and Quantity maxims (Grice 1967).

A central question that has been asked about pragmatic inferences such as scalar inference concerns their processing: does generating them incur processing cost? A large body of research has shown that implicature generation is indeed costly, as evidenced by increased reaction times (Bott & Noveck 2004), ERP data (Noveck & Posada 2003), or delays in eye-tracking (Huang & Snedeker 2009). These results support a literal-first model of implicature processing (Huang & Snedeker 2009). At the same time, however, some studies found evidence for a speedy implicature calculation process (Grodner et al. 2010) – a finding compatible with the default hypothesis about implicature processing (Levinson 2000). On the other hand, instead of making a categorical distinction between 'costly' vs. 'cost-free' processing, a constraint-based approach views implicature calculation and processing as resulting from the interaction of multiple cues and constraints (Degen & Tanenhaus 2015). Such a model privileges neither semantics nor pragmatics, and the research program is instead to identify and quantify the cues that modulate processing. In this paper we focus on one such cue: the Question Under Discussion (QUD).

In particular, we build on earlier work in the constraint-based framework by Degen (2013) and Degen & Tanenhaus (2015), which put forward the hypothesis that processing cost tracks the QUD. This predicts no uniform cost or lack of cost for calculating conversational implicatures per se. Instead, it attributes particular importance to supportive vs. non-supportive contexts, and predicts that the likelihood of inference calculation, and crucially also whether this calculation causes a delay in reaction times, is a function of how much the target inference is supported by the QUD. When a target inference is congruent with the QUD, an increased rate of inference generation is predicted, as well as decreased reaction times. When the target inference is not supported by the QUD, the opposite pattern is predicted. Our empirical findings confirm these predictions. Specifically, in Experiment 2 we show that under a QUD that makes the target inference likely to arise, inference computation does not cause a delay in reaction times. On the contrary, under a QUD that makes inference calculation unlikely, that calculation is time-consuming. Crucially, our QUD manipulation is grounded in an elicitation study (Experiment 1), which goes beyond earlier work by relying on experimental production data to establish likely QUDs, and thus constitutes a first attempt to address the problem of systematically identifying QUDs relevant to a given context.

The empirical case study for testing the QUD-sensitivity of conversational implicatures is first scalar inference, as exemplified in (1). Our finding that QUDs

modulate the calculation rate and reaction time cost of scalar inference is in line with existing work, successfully replicating previous results in a different experimental paradigm. We also extend these findings to a different kind of pragmatic inference, *it*-cleft exhaustivity (2):

(2) It is a cookie that Mary ate.
Literal: Mary ate a cookie and possibly other things too.
Inference-enriched: Mary **only** ate a cookie.

In the case of (2), just like in (1), hearers regularly go beyond the literal meaning of the sentence and calculate the inference-enriched meaning. This suggests that, at least on a descriptive level, *it*-cleft exhaustivity is similar to scalar inference. Our motivation for including *it*-cleft exhaustivity in our empirical scope is to probe whether the hypothesis that pragmatic inferences are QUD-sensitive is more widely applicable than just to *some but not all* inferences. There has also been recent interest in probing the extent to which different conversational implicatures, in our case scalar inference and *it*-cleft exhaustivity, pattern alike (see i.a. Chemla & Bott, 2014; van Tiel & Schaeken, 2017) [2].

We are aware that the question might be raised whether these two types of inference have the same theoretical status. For example, there exist different proposals in the literature about whether the exhaustivity effect in *it*-clefts arises as a conversational implicature (Horn, 1981; Drenhaus et al., 2011; Onea & Beaver, 2011; Byram Washburn et al., 2019), as a presupposition (Percus, 1997; Büring & Križ, 2013), or as part of the truth-conditional semantics (Atlas & Levinson 1981, É. Kiss 1998). Similarly, it has been argued that scalar inference is located in the grammar (and is computed locally, see i.a. Chierchia, 2004; Fox, 2007), or on the contrary, that it is in fact pragmatic and computed globally (Horn, 1972; Gazdar, 1979; Atlas & Levinson, 1981; Sauerland, 2004). Regardless of one's particular commitment about the status of these alternative-sensitive inferences, it is an empirical question whether computing inferences from these constructions is QUD-sensitive. In the current paper, we will thus sidestep the debate above and use the term *implicature* more broadly to refer to scalar inference and *it*-cleft exhaustivity. Our empirical findings showing that both scalar inference and *it*-cleft exhaustivity track QUDs can potentially inform that debate, but we leave that question for future work.

The rest of the paper is structured as follows. In Section 2.1, we discuss different hypotheses about the processing cost of implicatures, and in Section 2.2 the theoretical and experimental arguments for the importance of context. Section 2.3 outlines the goals and contributions of the present study. Section 3 reports on the QUD elicitation experiment we conducted, and Section 4 presents the findings of the QUD manipulation experiment. Section 5 discusses the broader implications of our findings. Section 6 concludes.

[2] The sentence in (2) also carries the existential presupposition that Mary ate something. We set this aside for the time being, but will return to it in our analysis of the experimental data.

2. BACKGROUND

In this section we first review different accounts of the processing of implicatures (2.1), followed by a discussion of previous work that investigated the influence of context and QUDs on the rate and cost of implicature calculation (2.2). Finally, we outline the contributions of the present paper: 1) an experimental elicitation paradigm probing what QUDs are relevant in a given context, and 2) testing the robustness of the QUD hypothesis using sentence-picture verification, extending its predictions to a previously untested pragmatic inference (2.3).

2.1. The cost of implicature processing

One of the major questions in psycholinguistic studies of semantics-pragmatics is how fast implicatures are processed: are they processed on par with other inferences, i.e. does generating them incur a cost? This is meant to inform our understanding of the relations between semantic and pragmatic processes during language comprehension. In the following we review three different approaches to this question.

2.1.1. Default hypothesis

Levinson (2000)'s default hypothesis takes implicatures to be default inferences, derived automatically and regardless of context. This predicts implicature calculation to be immediate and effortless, where inference-enriched interpretations will always precede and be accessed faster than literal interpretations. On the contrary, it is cancelling a conversational implicature that requires processing resources, and therefore time. This is offered as a solution to the puzzle of the communicative bottleneck: communication proceeds remarkably quickly despite physical limits on the rate of speech production. Experimental evidence for the default hypothesis comes i.a. from Grodner et al. (2010), who found eye-tracking evidence for a rapid interpretation of the inference-enriched meaning of *some*. Nonetheless, though Grodner et al. (2010) showed that scalar inferences are derived without processing delay, a further prediction of the default hypothesis, namely that the derivation of the literal interpretation (*some and possibly all*) should be associated with a processing delay, has not been confirmed.

2.1.2. Literal-first hypothesis

The literal-first hypothesis (put forward i.a. by Huang & Snedeker, 2009), assumes a two-stage processing sequence: literal interpretations are necessarily computed before, and therefore accessed faster than inference-enriched ones. Though scalar implicature processing may be rapid, it has to be preceded by some semantic analysis – in line with a model of linguistic architecture that takes semantic representations to serve as a mediator between phonology and pragmatics. A large body of research has indeed shown that implicature generation is costly, as evidenced by increased reaction times (Bott & Noveck

2004), ERP patterns (Noveck & Posada 2003), or delays in eye-tracking (Huang & Snedeker 2009).

2.1.3. Probabilistic frameworks

Under the default and literal-first hypotheses, semantics and pragmatics are clearly separated in the grammar and processing. Probabilistic frameworks, on the other hand, do not posit such a sharp boundary; rather, semantic and pragmatic processes are taken to be intertwined. Degen & Tanenhaus (2015)'s constraint-based framework (see also Degen & Tanenhaus, 2019 for an overview), for example, does not assign a privileged status to either literal or inference-enriched readings – neither reading is taken to always require more/less processing resources or time than the other. Instead, it assumed that how fast a scalar implicature is computed is in large part determined by context. Robustness of inference calculation, as well as speed and ease of processing, depend on the strength of cues available. The more probabilistic support there is from such cues, the faster comprehenders will arrive at the inference, and the less easily cancellable that inference will be.

The primary goal of these probabilistic models is not to test whether implicatures are calculated by default or at a cost, but rather to identify and quantify the cues that hearers rely on when generating inference-enriched meanings. Cues that have been shown to have an effect on the rate or cost of implicature calculation are e.g. the syntactic partitive, or the availability of lexical alternatives (Degen & Tanenhaus 2015); the relevance of the stronger alternative proposition (Breheny et al. 2006, Politzer-Ahles & Fiorentino 2013); cognitive load (De Neys & Schaeken 2007); the speaker's knowledge state (Goodman & Stuhlmüller 2013); or face threat (Bonnefon et al. 2009). QUDs are also taken to be such a probabilistic cue, and have thus been predicted to influence implicature calculation and processing (Degen 2013, Degen & Tanenhaus 2015). In the following, we turn to previous work on the role of context and QUDs.

2.2. The relevance of context

It has long been noted that, depending on the discourse context, otherwise predicted implicatures can fail to arise (see i.a. Kuppevelt, 1996). Following Roberts (1996/2012), we formalise context using the notion of Questions Under Discussion (QUDs), defined as the immediate topic of discussion, which proffer a set of relevant alternatives. Discourse is construed as giving rise to a stack of QUDs, and the ultimate discourse purpose is to answer all of these QUDs. An assertion is felicitous, then, if it chooses among the proffered alternatives and thereby bears upon the QUD. To see how QUDs could modulate implicature calculation, let's consider the example in (3):

(3) (a) Mary ate some of the cookies.
    (b) Mary ate all of the cookies.

(c) Mary ate some but not all of the cookies.

Assuming the QUD *How many cookies did Mary eat?*, *Mary ate some of the cookies* and *Mary ate all of the cookies* are both in the set of proffered alternatives. That is, both (3a) and (3b) would be felicitous responses to the QUD. The hearer of (3a) is therefore predicted to identify (3b) as a felicitous alternative, leading her to realise that the speaker's choice to utter (3a) implicates the negation of (3b), ultimately deriving the inference in (3c).

As the below example from Levinson (2000) shows, QUDs can also discourage the calculation of an implicature:

(4) A: Is there any evidence against them?
B: Some of their identity documents are forgeries.
Dispreferred implicature: Not all of their identity documents are forgeries.

The potentially available but dispreferred implicature in (4) would be consistent with the common ground and B's utterance; however, it is less likely to arise than the implicature in (3), because A's question suggests that she is only interested in whether there is at least some evidence against the criminals. Thus there is no particular reason to consider *All of their identity documents are forgeries* as an alternative that B could have said. We can thus see that QUDs offer a way to capture whether an implicature is more or less likely to arise[3].

As mentioned, under a probabilistic constraint-based model, QUD is one of many cues that is predicted to influence how likely an implicature is to be calculated. Predictions regarding the effect of QUDs on implicature calculation are supported by ample empirical data in experiments with both adults and children. It has long been observed that children are not adult-like in how they interpret structures that give rise to two or more competing readings, and are therefore more reluctant to calculate implicatures than adults (see e.g. Chierchia et al., 2001; Noveck, 2001). However, Papafragou & Tantalou (2004) showed that in contexts approximating naturalistic conversations, children can and do calculate implicatures. Investigating well-known lexical (*some-all*) as well as more context-dependent and ad hoc implicatures, the authors found that children exhibit robust rates of implicature calculation. Similar effects have been observed in the domain of scope ambiguities, where children are known to resort to surface (as opposed to inverse) scope interpretation more than adults (Musolino 1998, 2011), but actually show adult-like behaviour given the right context (Gualmini et al. 2008).

---

[3] The difference in implicature calculation between (3) and (4) may also be explained by appealing to the informativity of alternative answers. Wh-questions are commonly taken to carry an existential presupposition; here *How many cookies did Mary eat?* presupposes that there is some non-zero quantity of cookies that Mary ate. Given this QUD, the answer in (3a) is only informative on its inference-enriched interpretation. This contrasts with the polar-question QUD in (4), which carries no presupposition that there exists a non-zero quantity of evidence, making the answer informative on either its literal, or its inference-enriched reading.

QUDs have also been shown to induce variation in calculation rates in adult interpretations. Experiments have been carried out either using an explicit QUD (Zondervan et al., 2008; Yang et al., 2018), or promoting one implicitly via a background story (Degen 2013) or via focus intonation (Cummins & Rohde 2015). Zondervan et al. (2008), Degen (2013) and Yang et al. (2018) investigated *some but not all* scalar inferences, placing them under QUDs containing *some* vs. *all*, *none* vs. *all* and *any* vs. *all*, respectively. They found that reliably more inferences were calculated when an *all*-QUD is promoted. For example, Zondervan et al. (2008) presented participants with the sentence *Some pizzas were delivered*, which had to be evaluated with respect to either a question containing *some* (5), or one containing *all* (6).

(5) A: Were some pizzas delivered?
B: Some pizzas were delivered.

(6) A: Were all pizzas delivered?
B: Some pizzas were delivered.

The study found a 7% calculation rate of the *some but not all* inference in (5), but a 43% calculation rate in (6). Cummins & Rohde (2015) found comparable results testing a wider variety of scalar inferences, e.g. *warm-not hot*, which is especially important in light of recent findings that there is substantial variability among scales such as *some-all* vs. *warm-hot* with regard to the availability of the corresponding implicature (i.e. scalar diversity, van Tiel et al., 2016). This constitutes empirical confirmation that hearers do not always calculate scalar inference; rather, they are more or less likely to do so depending on context.

Constraint-based models predict QUDs to influence not only the likelihood of implicature calculation, but also the processing cost associated with this calculation. Indeed, Degen (2013) and Degen & Tanenhaus (2015) put forward the prediction that a QUD that makes the alternative *all* more relevant should lead to faster calculation of scalar inference than one that makes *none* salient. Degen (2013) explicitly manipulated the QUDs to test these predictions, and found that calculating the scalar inference is numerically faster under an *all*-QUD than under an *any*-QUD (Experiment 2a). Moreover, Degen (2013) and Degen & Tanenhaus (2015) also observed individual differences for scalar inference calculation, such that some participants consistently calculated the inference, some consistently did not, while a third group was inconsistent. The authors argued that participants' response consistency is indicative of how much uncertainty they had about the QUD. Inconsistent participants were argued to have more uncertainty, leading to a higher cost for generating the inference. Similarly, Kursat & Degen (2020) have found that reaction times are influenced by an interaction between the QUD and 'participant type' (i.e. whether or not a participant tends to calculate the inference), though their results did not reveal an overall modulating effect of QUDs.

2.3. Contributions of the present study

Existing work has thus hypothesised that the rate and speed of implicature calculation is not always uniform; rather, it depends on the QUD. Indeed, context has been shown to serve an important role in implicature calculation, modulating how likely an inference is to arise, as well as having an impact on response times. In this paper, our goal is twofold. First, we acknowledge the known problem that there exists no explicit mechanism for identifying the QUDs relevant for a given context. We make the first step to address this problem, and go beyond previous studies by establishing the relevant QUDs for a given utterance in a more empirically grounded manner, viz. by relying on experimental production data. The assumption that has often been made is that the relevant questions are the ones that contain a member of the given lexical scale, i.e. for the target sentence *Mary ate some of the cookies*, the possible questions would be *Did Mary eat some/none/all of the cookies?*. Although the assumption that these are the relevant QUDs is informed by theory, there has not been systematic empirical work probing the possible range of QUDs.

Second, we further test the hypothesis that pragmatic inference generation is QUD-sensitive (Degen & Tanenhaus 2015, Degen 2013). Specifically, we test two predictions of the QUD hypothesis. Our first prediction is that the effect of QUDs on calculation rates extends beyond the case of scalar inference to other types of quantity implicatures, specifically *it*-cleft exhaustivity. Second, and more importantly, we explore the prediction that QUDs modulate not only the calculation rates of implicatures, but also the reaction time cost of that calculation. The QUD hypothesis predicts that for both scalar inference and *it*-cleft exhaustivity, whether cost is incurred depends on the type of question the target sentence addresses. If a QUD supports the derivation of the inference (Inference-biasing), then there should be no increase in reaction times, but if the QUD biases against deriving the inference (Literal-biasing), processing cost is predicted.

Experiment 1 is the elicitation experiment aimed to empirically identify potential QUDs for a given context. Experiment 2 uses the elicited QUDs in a sentence-picture verification task, and shows that they modulate calculation rates and processing cost for both scalar inference and *it*-cleft exhaustivity. Additionally, we explore whether our findings might be informative not just regarding the presence or absence of processing cost for implicature calculation, but also regarding the exact source of this cost – especially as compared to the recent alternative-based Lexical Access account of van Tiel & Schaeken (2017) (Section 5.1).

3. EXPERIMENT 1: QUD ELICITATION

In this section we present an elicitation experiment, which established the likely QUDs for a given context for two types of quantity implicatures: scalar inference (SI) and *it*-cleft exhaustivity (EXH).

3.1. Participants

40 monolingual speakers of American English participated in an online elicitation experiment, administered on the Ibex platform (Drummond 2007). Participants were recruited on Amazon Mechanical Turk. Native speaker status was established via a language background questionnaire, where payment was not conditioned on the participant's response. Participants were compensated $2.00.

3.2. Task, materials and procedure

The task was a modified sentence-picture verification task used for elicitation. Participants were provided with a background story that two people, Anne and Bob, are discussing pictures about shapes. Anne cannot see these pictures, so she is always asking Bob about what he sees. The instructions also emphasised that Bob always answers truthfully. (7) shows the instructions given to participants before the start of the experiment.

(7) In this experiment you are going to see dialogues between Anne and Bob, who are discussing pictures. Each picture shows a number of colored shapes.
However, only Bob can see the pictures, Anne cannot. So Anne is always asking Bob about what he sees.

Participants then saw written SI and EXH target sentences paired with pictures, and were told that the sentences were Bob's answers to Anne's questions. Target sentences and pictures were adapted from the materials used by van Tiel & Schaeken (2017), with the addition of the context manipulation. The target sentences investigated were of the following form: SI: *Some of the shapes are blue* and EXH: *It is the square that is blue*. In place of Anne's question participants saw a blank line. We present an example of a trial screen in (8).

(8) Anne: _____*?*
Bob: *Some of the shapes are blue. / It is the square that is blue.*
Picture: Good Control or Target

The task was question elicitation: participants were instructed to guess what Anne's question could have been, based on Bob's answer and the picture, and had to provide their response in writing. We are aware that linguistically overt questions and the QUDs they evoke are different notions (see i.a. Hawkins et al., 2015 and references therein). Nevertheless, this task still gathers a measure of what participants think the target sentence is in response to, and is therefore a good proxy for tracking QUDs. For ease of explication, we will refer to the stimulus questions throughout as 'QUDs', which strictly speaking are the conversational topics tracked by the overt questions.

There were two types of pictures: Bob's answers were unambiguously good descriptions of Good Control pictures. For Target pictures, they were good descriptions on their literal (SI: *Some and possibly all of the shapes are blue*,

EXH: *The square is blue*), but not on their inference-enriched reading (SI: *Some but not all of the shapes are blue*, EXH: *Only the square is blue*). Examples of the pictures can be seen in Figures 1 and 2 for SI and EXH respectively.

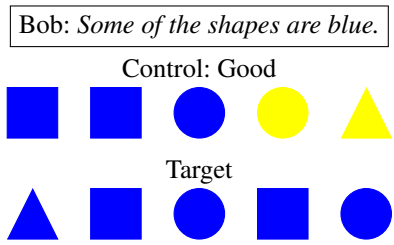

Figure 1
SI

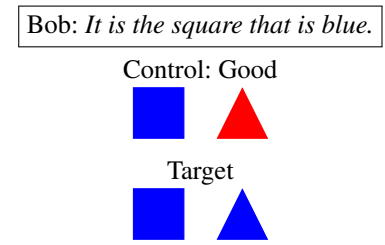

Figure 2
EXH

SI Good Control pictures always contained five shapes, three of which matched the colour mentioned in Bob's sentence (here, blue), while two were a different colour (here, yellow). SI Target pictures contained five shapes of the same colour, the one mentioned in the sentence. EXH Good Control pictures depicted the shape with the colour that was mentioned (here, blue square), as well as a different shape with a different colour (here, red triangle). EXH Target pictures depicted two shapes of the mentioned colour. Shapes were varied between triangle, circle and square (with SI pictures containing either a mix of shapes or the same shape five times), and colours were varied between blue, yellow, red, green, orange and black.

The experimental design included the two-level factor of Picture type as a between-participants manipulation: 20 participants saw only Good Control, and 20 other participants only Target pictures. A previous pilot with a within-participants design produced qualitatively the same results. Each participant saw 15 SI trials and 15 EXH trials. The experiment was administered in a Latin Square design.

3.3. Results and discussion

Results were coded in the following way. Whenever two responses only differed from each other in the mentioned colour or shape, they were coded as the same type of question. For example, *Are any of the triangles yellow?* and *Are any of the shapes red?* were both coded as *any*. Under each question type, we collapsed across closely related linguistic variants, for example *any of the shapes* and *any shapes* were both coded as *any*. In (9)-(10), we demonstrate the most commonly offered QUD types.

(9) Most frequent SI question types
what: *What color are the shapes?*
any: *Are any (of the) shapes blue? Are there (any) blue shapes?*

all: *Are all of the shapes blue?*
some: *Are some of the shapes blue?*

(10) Most frequent EXH question types:
which: *Which/what shape is blue? Which one (of them) is blue?*
any: *Are any of the shapes blue? Are there any blue shapes?*
what: *What color are the shapes? What color is the square?*

Though both SI and EXH elicited thirteen distinct types of questions each, the majority of answers came from a much smaller set for both constructions. Table 1 shows the frequencies of the most frequent types of questions in the data. The question types not included here occurred with less than 5% frequency[4].

	SI				EXH		
	what	any	all	some	which	any	what
Target	42%	25%	6%	12%	54%	9%	8%
Good Control	32%	33%	20%	2%	67%	14%	6%

Table 1
Frequencies of elicited question types

We can see that a small number of questions dominated in the elicitation task for each construction. But for SI, two types seem roughly equally frequent (*any*- and *what*-questions), whereas for EXH one type of question (*which*-questions) is clearly favoured over all others.

Another observation to be made about the data is that for both Good Control and Target pictures, the same type of questions were elicited and in largely the same frequencies. This is somewhat puzzling, considering that several of the elicited questions were previously thought to bias towards one or the other interpretation. For example, *all*-QUDs have been argued to bias towards the enriched reading, and *any*-QUDs towards the literal reading. Based on this argument, we might predict that only Good Control pictures, which are compatible with the inference-enriched meaning of SI, would have elicited e.g. *all*-questions. Conversely, we might not have predicted Target pictures, which are not compatible with the enriched reading, to elicit *all*-questions. One thing to keep in mind, however, is that in this experiment we do not obtain information about what interpretation participants actually assigned to the target sentences.

[4] The question types that each occurred with less than 5% frequency were the following. SI: *What's the dominant color?*, *Which color has the most shapes?*, *What color are some of the shapes?*, *What is one of the colors?*, *How many (shapes) are blue?*, *Which shapes are blue?*, *What is blue?*, *Are a lot of shapes blue?*, *If there are squares, what color are they?*; EXH: *Is one of the shapes blue?*, *What shape is it?*, *Which if any shapes are blue?*, *Is the triangle blue?* (7% frequency with Target), *Is it the circle or the square that is blue?*, *Are both shapes blue?*, *What is the color of the shape on the right?*, *Which shape is on the left?*, *What shape except for circle is blue?*, *Is it the square that is blue?*.

As many studies (including our Experiment 2) have found, some participants do judge *Some of the shapes are blue* to be a good description of both Good Control and Target pictures – in these cases, it is perhaps unsurprising that both pictures elicited the same type of questions. Another explanation might be that when participants provided *all*-questions in the Target condition, this was driven by the salience of the picture (where all shapes were blue), and not by question-answer congruence considerations. Future research should further probe these issues about the interplay of production and interpretation.

We also observe some differences between what kinds of questions the two construction types elicited, and in what frequency. As a reviewer points out, some of these differences may be due to the presuppositional difference between the SI and EXH constructions. We can see that SI sentences elicited more *what*-questions than EXH sentences, which in turn predominantly elicited *which*-questions. *Which shape is blue?* presupposes that there is a blue shape, but *What color are the shapes?* does not. The EXH construction similarly presupposes that there is something that is blue, while the SI construction asserts it. Therefore, participants may have supplied questions with the same presuppositional status as the target sentence, i.e. the answer: *which*-questions for EXH sentences and *what*-questions for SI sentences. The finding that *any*-questions were more frequent with SI than with EXH may also receive an explanation along these lines: *any*-questions are not presuppositional, but the EXH construction is. But although we recognize the appeal of this argument, the claim that question-answer pairs should be congruent in presuppositional status cannot account for the full set of data. In particular, according to such an argument, it is surprising that *any*-questions were elicited at all with EXH sentences: as we noted, the EXH construction carries a presupposition not present in the *any*-question; and given that the question-asker is stipulated to be ignorant, this presupposition should not already be in the common ground. But *any*-questions were in fact the second most frequent question type elicited given an EXH answer.

The primary aim of Experiment 1 was to test in a more systematic and empirically grounded manner what the likely QUDs are for a given dialogue context, where the dialogue includes either a *some* or an *it*-cleft sentence. Previous work on the role of context in SI calculation has largely assumed questions that contain members of the relevant lexical scale, while no investigation to our knowledge has been conducted about EXH in this respect. We can see that our elicitation study has indeed uncovered questions that have not been discussed in existing literature as relevant to implicature derivation or the particular target sentences, e.g. the *what*- and *which*-QUDs. On the other hand, some questions previously discussed in theoretical or experimental contexts did in fact show up in our elicitation data, e.g. *any*-, *some*- and *all*-QUDs.

Experiment 1 established likely QUDs whose effects on SI and EXH computation we can then investigate. We therefore took the most frequent questions for each construction from Experiment 1, and used them in Experiment 2 to see if they modulate implicature calculation rates and processing cost.

4. Experiment 2: QUDs modulate calculation rates and processing cost

We present a sentence-picture verification experiment, where we embedded scalar inference and *it*-cleft exhaustivity sentences under the most frequent QUDs elicited in Experiment 1. Our results showed that the probability of inference calculation, as well as whether there is an increase in reaction times, is conditioned on the QUD.

4.1. Participants

90 (30 in each QUD condition, different from those in Experiment 1) native monolingual speakers of American English participated in the experiment, administered on the Ibex platform (Drummond 2007). Participants were recruited on Amazon Mechanical Turk. Native speaker status was established via a language background questionnaire, where payment was not conditioned on the participant's response. Participants with a mistake rate exceeding 25% on filler items were removed from the analysis. This resulted in the removal of one participant from the wh-word experiment and four participants from the quantifier-experiment. Participants were compensated $1.50.

4.2. Task, materials and procedure

Experiment 2 employed the sentence-picture verification paradigm, with target sentences embedded in a dialogue context. Participants were given the same background story as in Experiment 1: Anne is asking questions from Bob, about pictures that only Bob can see. On the first screen, participants saw Anne's question, and they were instructed to press a key after they have read it. On the following screen, they saw Bob's answer to Anne's question, as well as the picture they were discussing. Participants were instructed to make a binary judgment (by clicking on a button) about whether Bob gave a good answer to Anne's question, given the picture he saw. The two buttons said 'Good' and 'Not Good'. Participants' choices were recorded, as well as reaction times from the onset of the second screen (displaying Bob's utterance and the picture) until the participant pressed one of the buttons.

The experiment featured a three-by-three design: crossing Picture (within-participants: Good Control, Bad Control, Target) and QUD (between-participants: wh-word, indefinite, quantifier). Similarly to Experiment 1, there were two types of pictures: Control, of which Bob's sentence was an unambiguously good/bad description, and Target, where the judgment depends on whether the inference has been derived – see Figures 3-4. Bob's answers are good descriptions of the Target pictures on their literal, but not on their inference-enriched reading. Good Control and Target pictures were identical to those used in the elicitation experiment. Note that for Experiment 2, Bad Controls were also added. SI Bad Control pictures contained five shapes, none of which has the colour (here, blue) mentioned in the

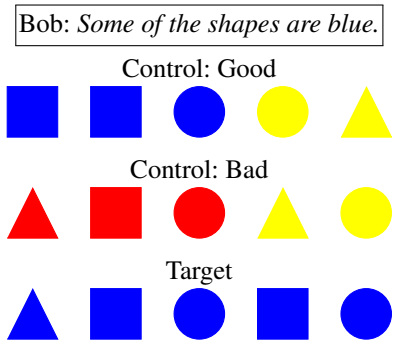

Figure 3
SI

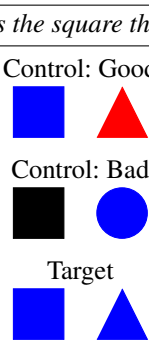

Figure 4
EXH

sentence. EXH Bad Controls contained the shape mentioned in the sentence, but in a colour different from the one in the sentence (here, black square), while the mentioned colour showed up on a different shape (here, blue circle).

The QUD condition (i.e. Anne's questions) was also a three-level manipulation: wh-word, indefinite, and quantifier questions. These questions were the most frequent elicited questions in Experiment 1 (see Table 1), with the exception of the *both*-question (in EXH), which we added as a counterpart to the *all*-question (in SI). (11)-(12) provide examples of the question manipulation for each construction. (Wherever there are two questions listed, half the participants saw one variant, and half the other variant.)

(11) QUD manipulation in SI (boldface for illustration only)
wh-word: **What** color are the shapes?
indefinite: Are there **any** blue shapes?/Are **any** shapes blue?
quantifier: Are **all** shapes blue?

(12) QUD manipulation in EXH (boldface for illustration only)
wh-word: **Which**/What shape is blue?
indefinite: Are there **any** blue shapes?
quantifier: Are **both** shapes blue?

QUD was manipulated between participants, placing each participant in either the wh-word, the indefinite, or the quantifier QUD condition. To prevent fatigue effects due to only encountering the same type of question, experimental items were intermixed with filler items, which included different shapes and questions unrelated to either inference, or QUD type. Each participant saw 12 SI, 12 EXH and 12 filler trials. The experiment was administered in a Latin Square design.

4.3. Predictions

Existing experimental work has shown that QUDs affect how likely an implicature is to be calculated (Section 2.2) – an effect we predict to extend to the previously

untested *it*-cleft exhaustivity. Based both on previous empirical results (Zondervan et al., 2008; Degen, 2013) and theoretical proposals (Hulsey et al., 2004), we can make the prediction that for SI, *all*-QUDs would result in the highest rate of implicature calculation. *Any*-QUDs, on the other hand (see e.g. (4)), are predicted to bias against deriving the implicature and lead to lower calculation rates. For EXH, parallel predictions can be made for the corresponding quantifier and indefinite: *both*-QUDs are predicted to lead to higher rates of implicature calculation than *any*-QUDs. The predictions for wh-question QUDs (SI *what*-QUD, EXH *which*-QUD) are less clear, in part because they have previously not been treated as relevant QUDs in these contexts, and it is less obvious what existing theoretical frameworks would predict about them (we discuss this in more detail in Section 5). We therefore treat their biasing behaviour as an empirical question. When analysing the results, our primary focus will be on the two other QUDs (*any* and *all* for SI; *any* and *both* for EXH), and our analysis of the wh-questions will be more exploratory.

Implicature calculation rates are indexed by the proportion of 'Not Good' responses to Target pictures: if a participant says that Bob gave a 'Not Good' description of a Target picture, she has calculated the SI/EXH inference. We thus predict variation across QUDs in the percentage of 'Not Good' responses to Target. For example, *any*-QUDs are predicted to result in lower rates of calculation and therefore lower 'Not Good' percentages for Target pictures. In what follows, we make a distinction between Literal-biasing QUDs, which lead to lower rates of inference generation (the prediction for e.g. *any*) and Inference-biasing QUDs, which lead to higher rates of inference generation (the prediction for e.g. *all*, *both*).

We consider longer RT when responding 'Not Good' to Target, relative to the RT when responding 'Not Good' to (Bad) Control, to be what indexes the cost of implicature calculation. This is because Bad Controls can be rejected based on literal, semantic meaning, but the rejection of a Target picture suggests that the participant has gone through the inference calculation process. In conducting such an analysis, we depart from Chemla & Bott (2014) and van Tiel & Schaeken (2017), who additionally analysed 'Good' responses and focused on the interaction of Condition and Response. Our reason for doing so is that responding 'Good' to Target pictures might indicate implicature non-calculation and reasoning only with the literal meaning, but it is also consistent with participants calculating, and then cancelling the implicature. Thus it is possible that for at least some participants or trials, responding 'Good' to Target may have included going through the inference calculation process. This would have led to increased RTs, introducing a confound for the interpretation. For this reason, in our analysis we focus on the Target vs. Control difference when responding 'Not Good', and disregard 'Good' responses to Targets.

Our predictions, then, are that under Inference-biasing QUDs (i.e. those that bias towards deriving an implicature) implicature derivation will not incur a cost. That is, we should not see a difference in 'Not Good' to Target as compared to

'Not Good' to Control RTs in Inference-biasing QUD conditions. On the other hand, Literal-biasing QUDs (i.e. those that bias against deriving an implicature) will have the effect that when the implicature is derived, its calculation is a costly and therefore slower process. Under Literal-biasing QUDs, therefore, we predict an increase in the time it takes to respond 'Not Good' to Target as compared to 'Not Good' to Control.

4.4. Results and analysis: rate of inference calculation

Prior to data analysis, we removed trials with extremely short or long response times by excluding the top and bottom 2,5% of the data based on reaction times.

Figures 5 and 6 plot the proportions of 'Not Good' responses for SI and EXH respectively. The primary purpose of Good and Bad Control pictures was to make sure participants are adequately doing the task, which we can see from responses being largely at floor and ceiling, respectively. Therefore, in the following, we focus on the analysis of the more informative Target pictures, which constitute our critical manipulation. Recall that for Target pictures, a 'Not Good' answer indicates implicature calculation. For the statistical analysis, a logistic regression model (glm function in R) was fit (mixed effects models did not converge), predicting Response (Good vs. Not Good) by QUD. Because our main prediction concerns the *any-all/both* difference, levels within the QUD variable were treatment coded, with *any* serving as the reference level. In SI, the statistical

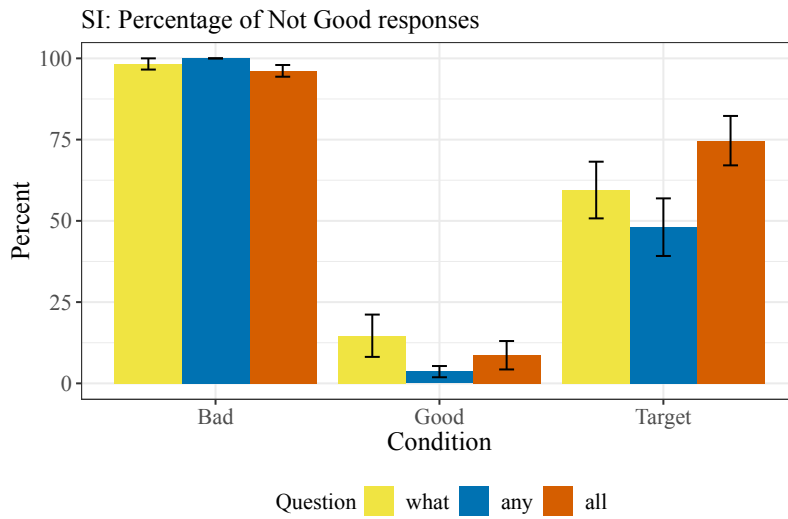

Figure 5
Proportion of participants' 'Not Good' (as opposed to 'Good') responses by Picture condition for *scalar inference*. Different colours denote the different QUD conditions. Error bars represent standard error.

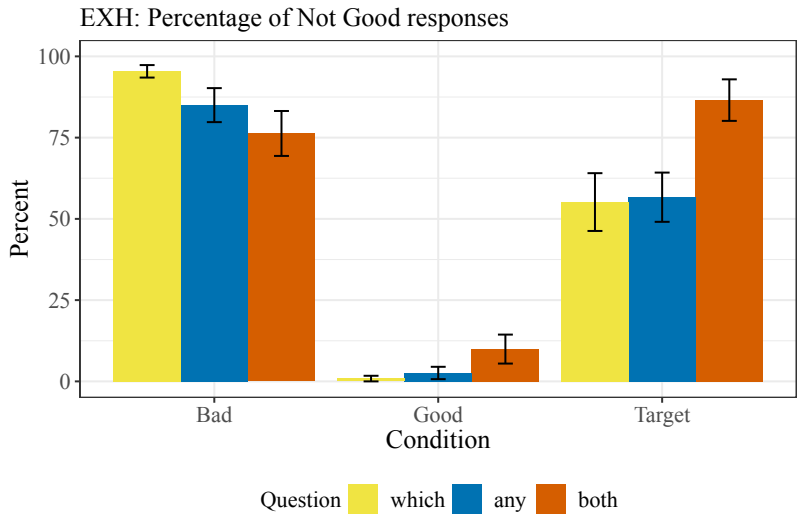

Figure 6
Proportion of participants' 'Not Good' (as opposed to 'Good') responses by Picture condition for *it-cleft exhaustivity*. Different colours denote the different QUD conditions. Error bars represent standard error.

analysis revealed that the difference between *all* vs. *any* ($p < 0.001$) is significant, while *what* resulted in responses somewhere in between the two other QUDs, with the *any* vs. *what* ($p < 0.06$) difference being only marginally significant (see Figure 5 and Table 2). An additional pair comparison between *all* and *what* revealed a marginally significant difference ($\beta = -0.58$, $z = -1.93$, $p < 0.06$). In EXH, *any* vs. *both* ($p < 0.001$) was revealed to be significantly different, but the *any* vs. *which* difference is not significant ($p = 0.7$) (see Figure 6 and Table 3). An additional pair comparison between *both* and *which* revealed a significant difference ($\beta = -1.64$, $z = -4.64$, $p < 0.001$).

Our prediction was that QUDs would modulate the rate of deriving pragmatic inferences, as indexed by rates of judging Bob's answer (as a description of

	Estimate	Std. Error	*z* value	*p* value
Intercept (any)	-0.06	0.19	-0.29	0.77
all	1.11	0.3	3.75	<0.001
what	0.53	0.28	1.92	0.05

Table 2
Scalar inference: Parameter estimates, standard errors, z values and p values from a logistic regression model of the 'Not Good' vs. 'Good' responses to Target, predicted by QUD.

	Estimate	Std. Error	*z* value	*p* value
Intercept (any)	0.33	0.19	1.71	0.09
both	1.55	0.36	4.36	<0.001
which	-0.09	0.27	-0.35	0.73

Table 3
It-cleft exhaustivity: Parameter estimates, standard errors, z values and p values from a logistic regression model of the 'Not Good' vs. 'Good' responses to Target, predicted by QUD.

the Target picture) 'Good' vs. 'Not Good'. Our findings are in line with these predictions. For SI, we found significantly fewer implicatures calculated under *any*-QUDs than under *all*-QUDs: that is, *any* is a Literal-biasing, while *all* is an Inference-biasing QUD. *What*-QUDs fall in the middle, making it unclear whether they can be categorised as either Literal- or Inference-biasing. In EXH, we also found significant differences in rate of implicature calculation, successfully extending earlier findings to a different kind of inference. Specifically, significantly fewer implicatures were calculated under *any*- and *which*-QUDs than under *both*-QUDs. In other words, for EXH, *any* and *which* are Literal-biasing, while *both* is an Inference-biasing QUD.

It is worth noting that under *any*-QUDs, there is a presuppositional mismatch between the EXH target sentence and the question: the EXH construction carries the existential presupposition that something is blue, while *any*-questions are not presuppositional. But it is instructive to note that in the Good Control condition, the rate of 'Good' responses was at ceiling. That is, participants almost always deemed the EXH sentence a 'Good' answer to an *any*-question, given a picture compatible with the exhaustivity inference. This suggests that participants were able to accommodate the existential presupposition and generate the exhaustivity inference, which is the phenomenon of interest to us.

Now that we have established that for both SI and EXH, some QUDs bias towards deriving the inference, while other QUDs bias against it, we turn to our main hypothesis. We show that these differences among QUDs are reflected in reaction time cost: the implicature calculation process will incur more or less reaction time cost depending on whether the target sentence is in a supportive context (i.e. under Inference-biasing QUDs) or a non-supportive context (i.e. under Literal-biasing QUDs). We analyse reaction time results in the next section.

4.5. Results and analysis: reaction times

Figures 7 and 8 plot reaction times broken down by QUD. Because our predictions concern the difference in RT when responding 'Not Good' to Target, as compared to Control (see Section 4.3), and we do not have specific hypotheses about differences in RTs when responding 'Good', we restrict the statistical analysis

to 'Not Good' responses. Nevertheless, the full data set is plotted in Figures 7-8. For the statistical analysis, a linear mixed effects model (lmer from the lme4 package in R, Bates et al., 2015) was fit, predicting RT by Condition (Target vs. Control). Levels within Condition were treatment coded, with Control serving as the reference level. Random slopes and intercepts were included for participants and items. Whenever the full model did not converge, the random effects structure was simplified following the recommendations of Barr et al. (2013). The p values reported below were estimated using the Satterthwaite procedure, as implemented in the lmerTest package in R (Kuznetsova et al. 2017). In the following we analyse RT data question-by-question, following the predictions we made.

4.5.1. Any-QUDs

Recall that *any*-QUDs were predicted to bias toward the literal meaning based on earlier theoretical and experimental work, and that is indeed what we found in our inference calculation rate data. Thus *any*-QUDs are predicted to make calculating an SI/EXH inference a costly process, manifested in longer RTs when responding 'Not Good' to Target as compared to Control. For the SI *any*-QUD, we found a significant difference in RT between responding 'Not Good' to Target vs. Control (β = 391.47, t = 2.6, $p < 0.05$). Similarly, for the EXH *any*-QUD, we found a significant difference in RT between responding 'Not Good' to Target vs. Control

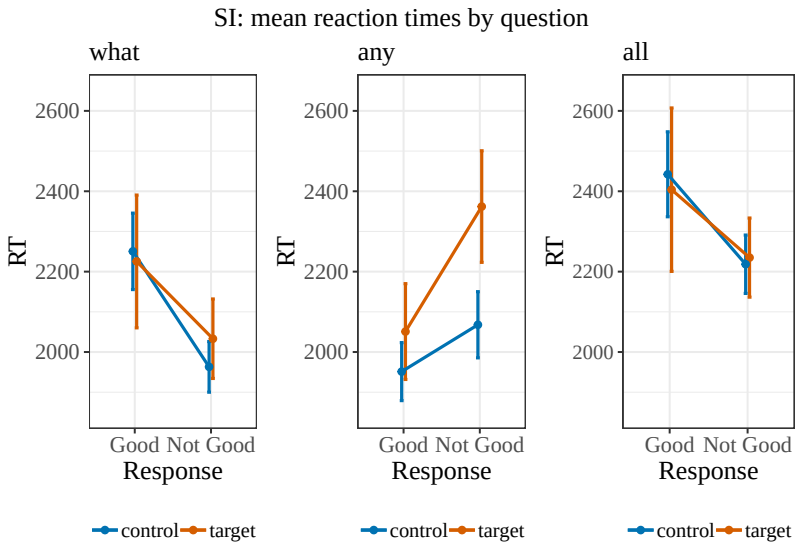

Figure 7
Mean reaction times (RT) in ms for judging Bob's answer as 'Good' or 'Not Good' in *scalar inference*. Different QUD conditions are displayed on separate plots. Colours denote Control (Good and Bad) vs. Target pictures. Error bars represent standard error.

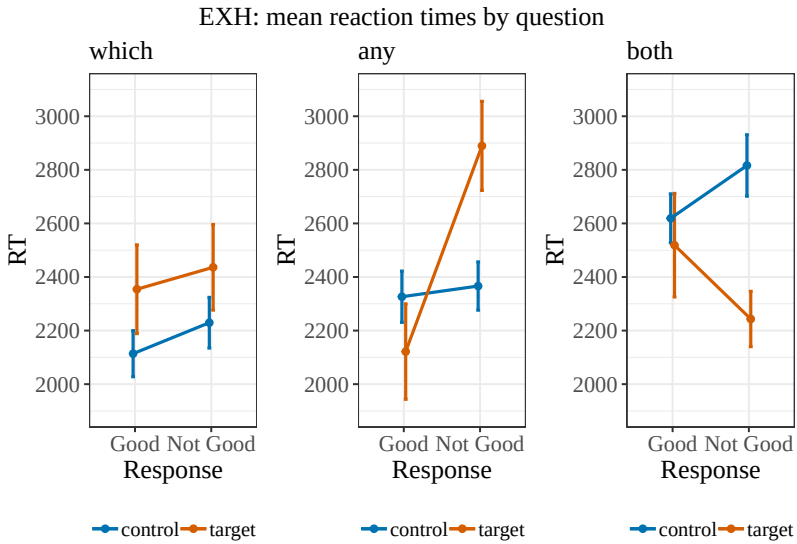

Figure 8
Mean reaction times (RT) in ms for judging Bob's answer as 'Good' or 'Not Good' in *it-cleft exhaustivity*. Different QUD conditions are displayed on separate plots. Colours denote Control (Good and Bad) vs. Target pictures. Error bars represent standard error.

(β = 539.01, t = 3.3, $p < 0.01$). Importantly, both patterns (see middle panels in Figures 7 and 8) show increased RT for the Target picture, as compared to the Control picture when responding 'Not Good'. This is thus in line with our prediction that *any*-QUDs make inference computation time-consuming.

4.5.2. All- and both-QUDs

The predictions made for the *all*- and *both*-QUD are the opposite of the prediction about the *any*-QUDs. *All*- and *both*-QUDs are Inference-biasing, and should therefore not lead to increased RTs when responding 'Not Good' to Target. For the SI *all*-QUD, there was no significant difference in RT between responding 'Not Good' to Target vs. Control (β = 5.06, t = 0.04, p = 0.97). For the EXH *both*-QUD, we found a significant difference in RT between responding 'Not Good' to the two types of pictures ($\beta = -478.01$, $t = -2.82$, $p < 0.01$), such that responding to Control resulted in longer RTs than responding to Target. That is (see the rightmost panel in Figure 8), while there was no reaction time cost for calculating the inference (i.e. no increased RTs for Target), we found an unexpected cost for responding 'Not Good' to the Control picture.

We argue that this unexpected cost is a side-effect of the picture stimuli. In particular, the verification process involved in rejecting a Bad Control involves two steps for EXH, but not for SI. For EXH (see Figure 10), given Bob's utterance

It is the square that is blue, both the colour and identity of each shape needs to be checked. This is because there is in fact only one blue shape (cf. exhaustive meaning) in the Bad Control picture, but that blue shape is not the correct one (i.e. the square). Note that all EXH stimuli necessarily have to include a blue shape, because the *it*-cleft sentence also carries an existential presupposition that something is blue. In contrast, for SI (see Figure 9), Bob's utterance *Some of the shapes are blue* can be rejected in one step, because no shapes in that display are blue.

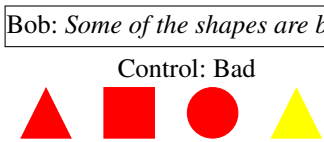

Figure 9
SI

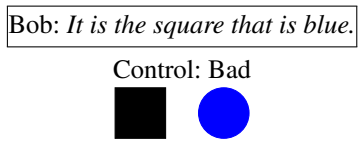

Figure 10
EXH

Nevertheless, our findings are generally in line with the predictions of the QUD hypothesis in that the Inference-biasing *all*- and *both*-QUDs did not lead to a delay in reaction times for deriving the SI/EXH inference.

Based on the above, we see a difference between *any*-QUDs, which led to processing cost, and *all*- and *both*-QUDs, which did not. To confirm that RT patterns vary across QUDs, we conducted an additional analysis focusing on the interaction of Condition with QUD. A linear mixed effects model (lmer) was fit, predicting RT by Condition (Target vs. Control), QUD (SI: *any* vs. *all*; EXH: *any* vs. *both*) and their interaction. Both variables were sum-coded, and models included random effects as described at the beginning of Section 4.5. For SI, we found a significant interaction of Condition with QUD ($\beta = 96$, $t = 2.1$, $p < 0.05$). Similarly, for EXH, we found a significant interaction of Condition with QUD ($\beta = 251.21$, $t = 4.62$, $p < 0.001$). That is, for both SI and EXH sentences, the RT patterns signalling processing cost were found to vary according to the preceding QUD.

4.5.3. What- and which-QUDs

Recall that we were not able to make clear predictions about *what*- and *which*-QUDs (Section 4.3), despite their frequency in the elicitation experiment. For the SI *what*-QUD, we found no significant difference in RT between responding 'Not Good' to Target vs. Control ($\beta = 194.91$, $t = 1.74$, $p = 0.1$). That is, the *what*-QUD parallels the *all*-QUD in resulting in no difference in RTs when responding 'Not Good' to Control vs. Target. This suggests that despite the mixed results in terms of SI calculation rate, the *what*-QUD is Inference-biasing based on reaction time measures. As for the EXH *which*-QUD, there was no significant difference in RT between responding 'Not Good' to Target vs. Control ($\beta = 266.6$, $t = 1.61$, $p = 0.11$). Qualitatively the *which*-QUD shows a somewhat similar pattern to the

any-QUDs, but the results are much less clear.

In sum, our predictions are largely borne out in the data: the questions which we could unambiguously classify as Literal- (*any*) vs. Inference-biasing (*all, both*) show divergent behaviour in terms of reaction time cost. *Any*-QUDs resulted in SI and EXH calculation being time-consuming, while the SI *all*-QUD and EXH *both*-QUD facilitated reaction times. The SI *what*-QUD patterned with Inference-biasing QUDs in that it did not lead to increased reaction times, while the EXH *which*-QUD was qualitatively similar to Literal-biasing *any*-QUDs. We discuss the implications of these results in the next section, along with the calculation rate results.

5. General discussion

In this paper we tested the hypothesis that the likelihood of calculating SI and EXH pragmatic inferences, as well as the processing cost of that calculation, tracks the QUD – a prediction of constraint-based models of implicature calculation. Under such a model, the more probabilistic support there is from multiple cues, the more quickly and robustly listeners will compute pragmatic inferences, with the QUD being one of the relevant cues. We elicited explicit questions to approximate the relevant QUDs (Experiment 1). Using these QUDs in Experiment 2, we found that they fall into one of two classes: under Literal-biasing QUDs, the rate at which participants drew inferences was lower, and under Inference-biasing QUDs, the rate at which they drew inferences was higher. Differences among QUDs also predicted processing cost. Under the Literal-biasing QUDs, making an inference-enriched judgment took longer than responding to the relevant literal control, whereas a facilitation of reaction time for such inferences was observed under Inference-biasing QUDs. By and large, these patterns hold for the majority of the questions we examined, although it is also worth noting that there appeared to be more nuanced patterns with wh-questions (see more discussion below). In general, we did not observe across-the-board processing cost for deriving implicatures, nor did we observe that computing inferences is always cost-free. Instead, our results strongly suggest that the cost of computing SI and EXH inferences is context-dependent – it is costly when the target expression was preceded by non-supportive QUDs. Such context-dependent cost imposes a significant constraint on our hypothesis space. For example, these results would be challenging for the *default hypothesis* and the *literal-first hypothesis* introduced in Section 2.1, since both would predict categorical behaviour that is QUD-independent.

Our findings contribute to the empirical landscape in a number of ways. First, we probed the robustness of earlier work on the QUD-sensitivity of scalar inference calculation and processing. Using a different experimental paradigm (sentence-picture verification) than existing work on QUDs, our findings successfully replicate the results of i.a. Zondervan et al. (2008), Degen (2013) and Cummins & Rohde (2015) (for calculation rates) and are in line with

trends observed by Degen (2013) and Degen & Tanenhaus (2015) (for reaction time). Moreover, this was done by manipulating the QUD directly via explicit questions, rather than implicitly via background stories – which the earlier processing experiments utilised. Crucially, we also found that QUD-sensitivity extends beyond the well-known case of scalar inference, to a previously untested pragmatic inference: *it*-cleft exhaustivity. Lastly, we took an important step towards better understanding how to empirically probe relevant QUDs, which we return to at the end of this section.

The specific role of QUDs observed in our study could be formalised under the Question-Answer Requirement (QAR, Hulsey et al., 2004) approach:

(13) *The Question Answer Requirement (QAR)*
The selected interpretation of an ambiguous sentence, whether true of false, is required to be a good answer to the Question Under Discussion. (A good answer is an interpretation that at least entails an answer to the QUD.)

In other words, any sentence is to be understood as an answer to a question, this question being the QUD (Hulsey et al., 2004; Gualmini et al., 2008). The QAR posits that the selected interpretation of an ambiguous sentence is required to be a good answer to the QUD. The two interpretations that SI target sentences allow for are repeated in (14) below.

(14) Some of the shapes are blue.
Literal: Some and possibly all of the shapes are blue.
Inference-enriched: Some but not all of the shapes are blue.

Using a standard Hamblin semantics for questions (Hamblin 1976), the SI results can be accommodated under the QAR as follows. On such a semantics, the meaning of a question, including the meaning of a QUD, is a partition of the set of possible worlds. The meaning of the QUD *Are there any blue shapes?* is a partition of worlds into the two sets (15a) and (15b):

(15) (a) $\{w : \exists x.x$ is a blue shape in $w\}$
(*Some and possibly all shapes are blue.*)
(b) $\{w : \neg\exists x.x$ is a blue shape in $w\}$
(*No shapes are blue.*)

Here, it is not necessary to derive the SI inference from *Some of the shapes are blue* in order to provide a good answer in the sense of the QAR, because the literal interpretation *Some and possibly all of the shapes are blue* corresponds to (15a). Deriving the inference also results in a good answer, because the inference-enriched interpretation (*Some but not all of the shapes are blue*) entails (15a). Therefore, under the *any*-QUD, the target sentence is a good response whether or not the inference is derived.

However, the picture changes for the *some but not all* SI target sentence under the QUD *Are all shapes blue?*, which partitions the set of worlds into the following two sets (assuming a five-shape display):

(16) (a) $\{w : \forall x.x$ is a shape $\rightarrow x$ is blue in $w\}$
(*All shapes are blue.*)
(b) $\{w : \neg\forall x.x$ is a shape $\rightarrow x$ is blue in $w\}$
(*Not all shapes are blue.*)

In this case, only if the inference is derived is the dialogue QAR-compliant, because the literal reading does not address the QUD. However, the enriched reading *Some but not all of the shapes are blue* entails (16b). Thus, the SI target sentence is only a good response to the *all*-question on the inference-enriched reading. This is reflected in our finding that participants derived significantly more implicatures with the *all*- than with the *any*-QUD, and that SI calculation led to a reaction time cost under the *any*-, but not under the *all*-QUD.

As for EXH, the two potential interpretation of the target sentence are repeated in (17).

(17) It is the square that is blue.
Literal: The square is blue.
Inference-enriched: Only the square is blue.

Assuming a two-shape display, the QAR captures the EXH findings in a way parallel to the SI findings. The EXH *any*- and *both*-QUDs result in the same partitioning of the set of worlds as the SI *any*- and *all*-QUDs respectively. The only difference is that the domain of quantification is now two instead of five, given the experimental pictures – as is reflected in the English paraphrases below.

(18) Partitioning from *any*-QUD (*Are there any blue shapes?*):
(a) $\{w : \exists x.x$ is a blue shape in $w\}$
(*One and possibly both shapes are blue.*)
(b) $\{w : \neg\exists x.x$ is a blue shape in $w\}$
(*Neither shape is blue.*)

(19) Partitioning from *both*-QUD (*Are both shapes blue?*):
(a) $\{w : \forall x.x$ is a shape$\rightarrow x$ is blue in $w\}$
(*Both shapes are blue.*)
(b) $\{w : \neg\forall x.x$ is a shape$\rightarrow x$ is blue in $w\}$
(*It is not the case that both shapes are blue.*)

Under the *any*-QUD, both the literal (*The square is blue)* and the inference-enriched (*Only the square is blue*) readings entail (18a). Therefore, both constitute good answers according to the QAR. However, under the *both*-QUD, only the inference-enriched reading answers the question by entailing (19b); the literal reading does not bear on the QUD. This difference is reflected in the empirical data: significantly more implicatures were derived under the *both*- than under the *any*-QUD, and only under the *any*-QUD was the computation time-consuming.

Additional to our main finding that context modulates the calculation and processing of SI and EXH, there remain some empirical puzzles regarding the

interplay of production and interpretation of QUDs. In Experiment 1, we took a first step in addressing the problem of narrowing down potential QUDs for a given context and conducted an elicitation study, the results of which fed into our QUD manipulation experiment. In doing so, we went beyond previous work that relied only on theoretically informed introspection to identify what may serve as a relevant QUD, and instead we treated this issue as an empirical question. Based on the results of the elicitation experiment, we focused on three types of questions: *any*-, *all*/*both*- and wh-QUDs. While the quantifier-QUDs have been discussed in existing literature as being relevant QUDs to the SI and EXH target sentences, the wh-QUDs (SI *what*-QUD, EXH *which*-QUD) have not. Yet it is interesting to observe that the novel wh-QUDs were the ones that proved most frequent in the elicitation experiment.

In addition to their novelty, the wh-QUDs also have some other puzzling properties. For instance, the focus structures of the *what*-QUD and the SI answer are potentially incongruent. The *what*-QUD focuses colors, yielding a set of alternatives of the form {*The shapes are red, The shapes are blue,* ...}; but one might think that the inference from the SI target sentence is most natural with focus on the quantifier *some* in *Some of the shapes are blue*, with an attendant set of focus alternatives {*None of the shapes are blue, Some of the shapes are blue, All of the shapes are blue*}. The wh-QUDs also showed more nuanced results in Experiment 2. The SI *what*-QUD did not fall neatly into the Literal- or Inference-biasing category based on the rate of inference calculation, though in RT results it qualitatively patterned with the SI *all*-QUD. As for the EXH *which*-QUD, while it showed distinct Inference-biasing behaviour in terms of likelihood of inference calculation, it resulted in mixed RT patterns. On the other hand, the QUDs that showed the clearest patterns (*any*-, *all*-QUDs) are in fact the ones with the closest link to existing theoretical work. Future work should thus address this tension between the observed mixed biasing and processing behaviour of *what*- and *which*-QUDs, and their apparent popularity with naive participants in an elicitation task.

The fact that elicitation resulted in somewhat surprising QUDs raises the issue of how to successfully probe theoretical constructs in an empirically grounded manner. One thing lacking in the current empirical picture is a measure of whether participants calculated the SI and EXH inferences in the elicitation experiment. Gathering such a measure could help address open questions regarding why the Target and Good Control conditions elicited very similar questions, as well as the surprising finding that wh-questions, which are not fully congruent with the target sentences in their presuppositions or focus structure, were also elicited. Conducting an elicitation experiment that also targets interpretation would thus be a valuable avenue for future work. Our study constitutes only a first step in understanding how QUDs can be elicited experimentally.

5.1. Lexical Access: alternative construction as the source of cost

Our findings highlight the role of QUDs as one source of the processing cost associated with implicature calculation. There are potentially other sources as well. Making reference to Katzir (2007)'s structurally based theory of alternative construction and complexity (see also Fox & Katzir, 2011), van Tiel & Schaeken (2017) (following Chemla & Bott, 2014) propose that the aspect of implicature calculation that causes a delay is lexical retrieval during alternative construction (Lexical Access hypothesis[5]). For example, in scalar inference, to construct the alternative *All of the shapes are blue* for *Some of the shapes are blue*, the lexical item *all* needs to be retrieved. The calculation of *it*-cleft exhaustivity, on the other hand, is argued to proceed without recourse to lexical alternatives. Van Tiel and Schaeken's (2017) empirical findings were in line with the Lexical Access hypothesis: it is only scalar inference (*some-all*) that resulted in processing effort, *it*-cleft exhaustivity did not.

The current findings pose an empirical challenge for the Lexical Access account. While evidence for Lexical Access comes from studies that presented target sentences in the absence of any context, our experiments varied the QUD. Our results do not support the prediction that the calculation of SI inferences, but not the calculation of EXH inferences, would lead to processing cost. Instead, we found that depending on context, both inferences can lead to a reaction time delay (or the lack of a delay). This effect of context-dependence is unexpected if the cost of calculating pragmatic inferences is directly and uniquely tied to the construction and complexity of the relevant alternatives.

Particularly informative in this respect are our findings about EXH, as well as SI *what*-QUDs. The predictions of Lexical Access converge with that of a QUD-based hypothesis about processing cost in the case of questions that explicitly mention alternatives, e.g. *all*-QUDs. Following a context (in our study: an explicit question) where the alternative *all* has been made salient, lexical retrieval of *all* will likely be a faster process. However, it is less clear how the Lexical Access hypothesis would capture our findings about questions that do not explicitly mention relevant lexical alternatives: the *what*-QUD, which showed reaction time patterns similar to the *all*-QUD, or the findings about EXH, where no lexical alternatives are relevant.

It is possible that both lexical retrieval and context-dependence contribute to the complexity of generating pragmatic inferences. Our results do not necessarily rule out that lexical access plays a role. Future research should probe further whether these diverse sources of complexity could selectively target different aspects of the processing cost generated during the inferential process.

---

[5] Lexical Access is of great contemporary interest not only as a proposed source of processing cost when calculating pragmatic inferences, but also to capture child language data. In particular, recent work has argued that children's divergence from adult behaviour when it comes to the interpretation of scalar inference or disjunction is due to their inability to access lexical alternatives (see i.a. Barner et al., 2011; Singh et al., 2016)

6. CONCLUSION

How people generate pragmatic implicatures has been a central question in linguistics, as it serves as a window into the process that integrates semantic and pragmatic information. The current study found that QUDs affect the likelihood of inference calculation. Our empirical case studies included both scalar inferences and *it*-cleft exhaustivity. Also importantly, for both kinds of pragmatic inferences, QUDs also predicted the amplitude of the processing cost (measured by reaction time). Altogether, our findings provide support for the constraint-based framework, and also open up questions about the precise source(s) of the complexity associated with computing pragmatic inferences.

REFERENCES

Atlas, Jay D. & Stephen C. Levinson. 1981. It-clefts, informativeness and logical form. In P. Cole (ed.), *Radical pragmatics*, 1–62. New York: Academic Press.

Barner, David, Neon Brooks & Alan Bale. 2011. Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition* 118(1). 84–93. doi:10.1016/j.cognition.2010.10.010.

Barr, Dale J, Roger Levy, Christoph Scheepers & Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3). 255–278. doi:10.1016/j.jml.2012.11.001.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. doi:10.18637/jss.v067.i01.

Bonnefon, Jean-François, Aidan Feeney & Gaëlle Villejoubert. 2009. When some is actually all: Scalar inferences in face-threatening contexts. *Cognition* 112(2). 249–258. doi:10.1016/j.cognition.2009.05.005.

Bott, Lewis & Ira A. Noveck. 2004. Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language* 51(3). 437–457. doi:10.1016/j.jml.2004.05.006.

Breheny, Richard, Napoleon Katsos & John Williams. 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. doi:10.1016/j.cognition.2005.07.003.

Büring, Daniel & Manuel Križ. 2013. It's that, and that's it! Exhaustivity and homogeneity presuppositions in clefts (and definites). *Semantics and Pragmatics* 6(6). 1–29. doi:10.3765/sp.6.6.

Byram Washburn, Mary, Elsi Kaiser & Maria Luisa Zubizarreta. 2019. *The English it-cleft: No need to get exhausted* 198–236. Leiden, The Netherlands: Brill. doi:10.1163/9789004378308_006.

Chemla, Emmanuel & Lewis Bott. 2014. Processing inferences at the semantics/pragmatics frontier: disjunctions and free choice. *Cognition* 130(3). 380–396. doi:10.1016/j.cognition.2013.11.013.

Chierchia, Gennaro. 2004. Scalar implicatures, polarity phenomena and the syntax/pragmatics interface. In Adriana Belletti (ed.), *Structures and beyond*, 39–103. Oxford University Press.

Chierchia, Gennaro, Stephen Crain, Maria Teresa Guasti, Andrea Gualmini & Luisa Meroni. 2001. The acquisition of disjunction: Evidence for a grammatical view of scalar implicatures. In Aimee Johansen Anna H.-J. Do, Laura Domínguez (ed.), *BUCLD 25 Proceedings*, Somerville, Massachusetts: Cascadilla Press.

Cummins, Chris & Hannah Rohde. 2015. Evoking context with contrastive stress: Effects on pragmatic enrichment. *Frontiers in Psychology* 6. 1779. doi:10.3389/fpsyg.2015.01779.

De Neys, Wim & Walter Schaeken. 2007. When people are more logical under cognitive load – Dual task impact on scalar implicature. *Experimental Psychology* 54(2). 128–133. doi:10.1027/1618-3169.54.2.128.

Degen, Judith. 2013. *Alternatives in pragmatic reasoning*: University of Rochester dissertation.

Degen, Judith & Michael K. Tanenhaus. 2015. Processing scalar implicature: A constraint-based approach. *Cognitive Science* 39(4). 667–710. doi:10.1111/cogs.12171.

Degen, Judith & Michael K. Tanenhaus. 2019. Constraint-based pragmatic processing. In Chris Cummins & Napoleon Katsos (eds.), *Handbook of experimental semantics and pragmatics*, Oxford: Oxford University Press.

Drenhaus, Heiner, Malte Zimmermann & Shravan Vasishth. 2011. Exhaustiveness effects in clefts are not truth-functional. *Journal of Neurolinguistics* 24(3). 320–337. doi:10.1016/j.jneuroling.2010.10.004.

Drummond, Alex. 2007. Ibex Farm. http://spellout.net/ibexfarm.

É. Kiss, Katalin. 1998. Identificational focus versus information focus. *Language* 74(2). 245–273. doi:10.1353/lan.1998.0211.

Fox, Danny. 2007. Free choice and the theory of scalar implicatures. In Uli Sauerland & Penka Stateva (eds.), *Presupposition and implicature in compositional semantics*, 71–120. London: Palgrave Macmillan UK. doi:10.1057/9780230210752_4.

Fox, Danny & Roni Katzir. 2011. On the characterization of alternatives. *Natural Language Semantics* 19(1). 87–107. doi:10.1007/s11050-010-9065-3.

Gazdar, Gerald. 1979. *Pragmatics: Implicature, presupposition and logical form*. New York: Academic Press.

Goodman, Noah D. & Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science* 5(1). 173–184. doi:10.1111/tops.12007.

Grice, Herbert Paul. 1967. Logic and conversation. In Paul Grice (ed.), *Studies in the way of words*, 41–58. Harvard University Press.

Grodner, Daniel J., Natalie M. Klein, Kathleen M. Carbary & Michael K. Tanenhaus. 2010. "Some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition* 116(1). 42–55. doi:10.1016/j.cognition.2010.03.014.

Gualmini, Andrea, Sarah Hulsey, Valentine Hacquard & Danny Fox. 2008. The Question-Answer Requirement for scope assignment. *Natural Language Semantics* 16(3). 205–237. doi:10.1007/s11050-008-9029-z.

Hamblin, Charles Leonard. 1976. Questions in Montague English. In Barbara H. Partee (ed.), *Montague grammar*, 247–259. Academic Press. doi:10.1016/B978-0-12-545850-4.50014-5.

Hawkins, Robert X. D., Andreas Stuhlmüller, Judith Degen & Noah D. Goodman. 2015. Why do you ask? Good questions provoke informative answers. In David C. Noelle, Rick Dale, Anne Warlaumont, Jeff Yoshimi, Teenie Matlock, Carolyn Jennings & Paul P. Maglio (eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, 878–883. Austin, TX: Cognitive Science Society.

Horn, Lawrence R. 1972. *On the semantic properties of the logical operators in English*: UCLA dissertation.

Horn, Lawrence R. 1981. Exhaustiveness and the semantics of clefts. In Victoria Burke & James Pustejovsky (eds.), *Proceedings of the North East Linguistic Society (NELS)*, vol. 11, 125–142.

Huang, Yi Ting & Jesse Snedeker. 2009. Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive Psychology* 58(3). 376–415. doi:10.1016/j.cogpsych.2008.09.001.

Hulsey, Sarah, Valentine Hacquard, Danny Fox & Andrea Gualmini. 2004. The Question-Answer Requirement and scope assignment. In Aniko Csirmaz, Andrea Gualmini & Andrew Nevins (eds.), *MIT Working Papers in Linguistics*, 71–90. MITWPL.

Katzir, Roni. 2007. Structurally-defined alternatives. *Linguistics and Philosophy* 30(6). 669–690. doi:10.1007/s10988-008-9029-y.

Kuppevelt, Jan van. 1996. Inferring from topics: scalar implicatures as topic-dependent inferences. *Linguistics and Philosophy* 19(4). 393–443.

Kursat, Leyla & Judith Degen. 2020. Probability and processing speed of scalar inferences is context-dependent. In Stephanie Denison, Michael Mack, Yang Xu & Blair C. Armstrong (eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, 1236–1242. Cognitive Science Society.

Kuznetsova, Alexandra, Per B. Brockhoff & Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82(13). 1–26. doi:10.18637/jss.v082.i13.

Levinson, Stephen C. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. The MIT Press.

Musolino, Julien. 1998. *Universal grammar and the acquisition of semantic knowledge: An experimental investigation of quantifier–negation interaction in English*: University of Maryland, College Park dissertation.

Musolino, Julien. 2011. Studying language acquisition through the prism of isomorphism. In Jill de Villiers & Tom Roeper (eds.), *Handbook of generative approaches to language acquisition*, 319–349. Dordrecht: Springer Netherlands. doi:10.1007/978-94-007-1688-9.

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

Noveck, Ira A. & Andres Posada. 2003. Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language* 85(2). 203–210. doi:10.1016/S0093-934X(03)00053-1.

Onea, Edgar & David Beaver. 2011. Hungarian focus is not exhausted. In Ed Cormany, Satoshi Ito & David Lutz (eds.), *Proceedings of Semantics and Linguistic Theory (SALT) 19*, 342–359. doi: 10.3765/salt.v19i0.2524.

Papafragou, Anna & Niki Tantalou. 2004. Children's computation of implicatures. *Language Acquisition* 12(1). 71–82.

Percus, Orin. 1997. Prying open the cleft. In *Proceedings of the North East Linguistic Society (NELS)*, vol. 27, 337–351.

Politzer-Ahles, Stephen & Robert Fiorentino. 2013. The realization of scalar inferences: Context sensitivity without processing cost. *PLOS ONE* 8(5). 1–6. doi:10.1371/journal.pone.0063943.

Roberts, Craige. 1996/2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics* 5(6). 1–69. doi:10.3765/sp.5.6.

Sauerland, Uli. 2004. Scalar implicatures in complex sentences. *Linguistics and Philosophy* 27(3). 367–391. doi:10.1023/B:LING.0000023378.71748.db.

Singh, Raj, Ken Wexler, Andrea Astle-Rahim, Deepthi Kamawar & Danny Fox. 2016. Children interpret disjunction as conjunction: Consequences for theories of implicature and child development. *Natural Language Semantics* 24(4). 305–352. doi:10.1007/s11050-016-9126-3.

van Tiel, Bob, Emiel Van Miltenburg, Natalia Zevakhina & Bart Geurts. 2016. Scalar diversity. *Journal of Semantics* 33(1). 137–175. doi:10.1093/jos/ffu017.

van Tiel, Bob & Walter Schaeken. 2017. Processing conversational implicatures: alternatives and counterfactual reasoning. *Cognitive Science* 41. 1119–1154. doi:10.1111/cogs.12362.

Yang, Xiao, Utako Minai & Robert Fiorentino. 2018. Context-sensitivity and individual differences in the derivation of scalar implicature. *Frontiers in Psychology* 9. 1720. doi:10.3389/fpsyg.2018.01720.

Zondervan, Arjen, Luisa Meroni & Andrea Gualmini. 2008. Experiments on the role of the question under discussion for ambiguity resolution and implicature computation in adults. In Tova Friedman & Satoshi Ito (eds.), *Proceedings of Semantics and Linguistic Theory (SALT) 18*, 765–777. doi: 10.3765/salt.v18i0.2486.

Authors' addresses: *(Ronai)*
The University of Chicago, 1115 E. 58th Street, Chicago, IL 60637, USA
ronai@uchicago.edu

(Xiang)
The University of Chicago, 1115 E. 58th Street, Chicago, IL 60637, USA
mxiang@uchicago.edu