Contextual causes of implicature failure

Chris Cummins

Department of Linguistics and English Language, University of Edinburgh

Abstract

Theoretical and empirical research on quantity implicature has concurred that pragmatically strengthened, richer readings are not available when they are not relevant to the discourse purpose. However, this claim relies on an appeal to a notion of "relevance" which has proved difficult to make precise. In this paper I discuss and contrast two potential contributory factors to relevance: adherence to the QUD, and form-based priming effects. The former can be considered to operate at a relatively high level of analysis, from the speaker's perspective, and influences the semantic content that the speaker should be attempting to convey, while the latter is assumed to reflect low-level psychological preferences and influences the form of words that the speaker should use. I argue that pragmatics, and specifically implicature, constitutes a useful testbed for distinguishing these effects - the availability of an implicature can be used as an indicator that a particular stronger alternative would also have been an acceptable utterance, while its unavailability suggests that the chosen utterance is preferable to this stronger alternative for one reason or another. I discuss recent experimental data from this perspective, and argue that both QUD and priming effects are customarily at play. I conclude by exploring the implications of this for our view of pragmatics and its interfaces.

Introduction

The nature of quantity implicature is a much-discussed question in the recent pragmatics literature. One particular focus of attention is whether implicatures are recovered in a traditionally pragmatic fashion (for instance, in accordance with a Gricean account) or whether they are obtained as default inferences (Levinson 2000) or via a grammatical process of parsing (Chierchia 2006). A substantial volume of experimental research has been brought to bear on this issue over the past few years, but arguably the results have been inconclusive.

Work on the time-course of processing has found little evidence for default inferences, but has relied upon the assumption that such default processing would necessarily be less effort (and therefore faster) than Gricean-style reasoning. More recently, work on so-called embedded implicatures, which are predicted to be generally unavailable under a Gricean account, has sparked a fierce debate as to the correct theoretical interpretation of the experimental findings (Geurts and Pouscoulous 2009, Chemla and Spector 2011).

Although theoretically inconclusive, this body of experimental data has substantially informed our understanding of the factors bearing upon the ultimate interpretation of potential implicature triggers. For instance, Goodman and Stuhlmüller (2013) (see also Cummins and Katsos 2010) show that hearers factor in the speaker's knowledge state, and do not recover quantity implicatures (e.g. interpreting "some" as "not all") in cases where the speaker is not knowledgeable about the stronger proposition (in this example "all"). Bonnefon, Feeney and Villejoubert (2011) show that hearers also refrain from recovering such implicatures in cases where the stronger proposition would have been face-threatening. Most pertinently for the purposes of this paper, Breheny, Katsos and Williams (2006) demonstrated that participants did not exhibit implicatures in cases where the stronger alternative would not have been necessary given the purpose of the discourse.

These results are largely predictable under a classical pragmatic account of quantity implicature (although they can also be encompassed by the alternative accounts, given auxiliary assumptions). On the traditional view, a quantity implicature is argued to arise as a consequence of the speaker electing to make a particular statement instead of an informationally stronger one. Consequently, no implicature is predicted to arise in cases where the speaker is effectively prohibited from making a stronger statement by external factors. If the speaker does not know whether or not the stronger statement is true, the social obligation to avoid making potentially false statements (couched by Grice (1975) as the maxim of quality) prevents them from asserting it. Similarly, if the stronger statement would be face-threatening, we might reasonably suppose social considerations to militate against its use (as codified, for instance, by Leech's (1983) maxim of Approbation). In both cases, the speaker does not have a free hand to select an utterance purely on the basis of its informational value, so the preconditions for quantity implicature are not satisfied.

The third case, in which the stronger statement is not necessary for the discourse purpose, may also be explicable in similar terms, but is perhaps a subtler matter. To illustrate this, consider (1) and (2), sample materials from Breheny et al. (2006).

(1) Mary asked John whether he intended to host all his relatives in his tiny apartment. John replied that he intended to host some of his relatives. The rest would stay in a nearby hotel.

(2) Mary was surprised to see John cleaning his apartment and she asked the reason why. John replied that he intended to host some of his relatives. The rest would stay in a nearby hotel.

The premise of this pair of examples is that (1) presents its second sentence, which contains the potential implicature trigger "some", in an upper-bound context – that is, one in which the stronger alternative "all" would be relevant to the discourse purpose (answering Mary's question). By contrast, (2) presents its second sentence in a lower-bound context, in which the alternative "all" would not be relevant – the fact that John is hosting some of his relatives already suffices to explain why he is cleaning his apartment. In a self-paced reading experiment, Breheny et al. (2006) document that reading times for "the rest" are faster in items such as (1) than in items such as (2), suggesting that readers of (1) are conscious that the enriched reading of “some” (“some but not all”) is intended here (and consequently are aware of the existence of a complement set that can be referred to by “the rest”).

In this paper, I aim to look more closely at instances of implicature suppression that could theoretically be attributed to the lack of relevance of the stronger proposition. I will suggest that these effects might typically be traceable either to high-level considerations of discourse organisation, such as can be expressed in terms of QUD, or to low-level mechanistic factors such as priming. I consider the extent to which these factors can be disentangled, and what the availability of implicature might tell us about the way in which QUD is constituted.

Suppressing implicatures on relevance grounds

The preferred account of how the implicature “not all” arises in cases such as (2) depends on the theoretical framework that we wish to adopt. From a Gricean perspective, we can think of the use of “some” in (2) – unlike in (1) – as a possible consequence of the second submaxim of quantity, “Do not make your contribution more informative than is required”. We could argue that, for John to announce that he intended to host all of his relatives – even if this was in fact the case – would potentially violate this submaxim. Consequently, as hearers, we would be entitled to surmise that the use of “some” might merely reflect an unwillingness to be overinformative and use “all”, rather than an attempt to convey that “all” was not the case. For this reason, we would again be able to predict the suppression of the implicature, on the basis that the use of “some” is not a reflection of the speaker’s choice but merely due to the obligation to obey the second submaxim of quantity.

This argument parallels the account given in the previous section for the epistemically-driven suppression of implicature – that is, the tendency for “some” not to implicate “not all” when the speaker is not knowledgeable about the proposition with “all”.

However, one problem with extending that argument in this way is that there is a substantial asymmetry in strength between the maxim of quality and the second submaxim of quantity. Grice (1975: 46) felt that the former was a *sine qua non* of communication to the extent that it might not belong alongside the other maxims, which "come into operation only on the assumption that this maxim of Quality is satisfied". By contrast, overinformativeness – while generally dispreferred – does not seem to be a serious barrier to communication (see Davies and Katsos 2011 for relevant experimental data). Moreover, it seems, on the face of it, completely reasonable and quite probable that a speaker would use "all" in (2) if they knew the resulting statement to be true. Hence, the argument that there is no implicature because the speaker is prohibited from using the stronger alternative, "all", does not seem entirely tenable in this case.

As a consequence, it is tempting to suppose that the lack of spontaneous inference from "some" to "not all" in (2) might be due to the reader adopting a particular pragmatic strategy: for instance, stopping enrichment when a particular level of relevance is achieved (in a relevance-theoretic sense), or looking for the simplest answer to the Question Under Discussion (QUD).

From a Relevance Theory (RT) perspective, it is predicted that hearers will perform pragmatic enrichments in such a way as to optimise relevance, where this is construed as the ratio of cognitive effects to cognitive effort. From this perspective, we can identify (1) as a case in which the enrichment from "some" to "some but not all" results in optimal relevance being achieved and (2) as a case in which it does not. One limitation of this approach, which seems to be shared by experimental literature in an RT framework (for instance, Van der Henst, Carles and Sperber 2002), is that the precise notion of cognitive effects and cognitive effort are not really quantified, and consequently it is not entirely clear whether the predictions of RT are falsifiable. By common consent, enriching "some" in (2) would still result in additional cognitive effects, namely the hearer internalising the truth of the proposition that “John does not intend to host all of his relatives”. On the basis that this does not happen, we conclude – given the assumptions of RT – that these effects are not sufficient to counterweigh the additional effort involved in computing this enrichment. In this way, the data can be accommodated within the theory, but the theory offers no *a priori* way of predicting whether or not this enrichment will take place.

Within such a framework, it appears that we would also need to stipulate some way in which a hearer can evaluate whether the enrichment would be useful enough to make it worth obtaining, without actually calculating the enrichment and exploring its consequences (because once the hearer has expended the cognitive effort of doing this, the most ‘relevance-enhancing’ thing to do would be also to take advantage of the cognitive effects that go with it). That is, to know that it is not advantageous to enrich “some” in (2), the hearer must know that the possible enrichment is “some but not all” and that that enrichment does not contain additional useful information – or at least not to the extent that would make it worth computing. But working out what the implicature would be, in order to reject it, clearly results in a less favourable outcome relevance-wise than simply ignoring the possibility of the implicature outright: this latter option incurs less cognitive costs and results in the same cognitive effects. Intuitively, it certainly appears that we do not tend to perform complex pragmatic computations about irrelevant stronger propositions, which suggests that we have some form of heuristic for excluding them from consideration.

The idea of QUD offers one way of fleshing out this intuition. We can distinguish cases such as (1) and (2) by noting that the sentence with “some” is directed towards a different QUD in the two cases. In the former, the QUD appears to be “whether John will host all of his relatives”, and in the latter, it appears to be “why John is cleaning his apartment”. In the former case, but not the latter, the enrichment from “some” to “some but not all" is required in order for John's reported utterance to qualify as an answer to this QUD. In particular, this would offer us a convenient way of determining when to stop enriching the utterance pragmatically, namely that we do so when we have satisfactorily divined the answer to the present QUD. Of course, in this particular experimental example, the QUD is made highly explicit in the prior discourse context: as Carlson (1982) observes (cited in Roberts 2012: 8), though dialogues may be ordered by question-answer relations, the questions themselves are often only implicit and must be inferred on the basis of other cues.

At a somewhat lower level of organisation, we could argue that there is a distinction between (1) and (2) in that (1) involves explicit prior mention of the term "all (of his relatives)", against which "some (of his relatives)" could be considered contrastive. As argued by Geurts and van Tiel (2013), contrastive environments of this kind are widely acknowledged to give rise to truth-conditional narrowing effects, which need not necessarily be analysable as implicatures in the Gricean sense. Thus, we could argue that "some" in (1) may convey "not all" simply by virtue of the fact that it is not the word "all" that is used. In (2), however, no such contrast exists, and no such enrichment proceeds.

We could even go further and speculate that the pragmatic difference between (1) and (2) stems from considerations of priming. In (1), the sentence structure from Mary's question is largely repeated in John's response ("…he intended to host…"). If we assume that John is bound by the kind of syntactic priming effect discussed by Branigan et al. (2000) and Pickering and Garrod (2004), we might have a strong expectation that, if "all" were the case, then he would use the primed form of words ("he intended to host all his relatives"). Consequently, John's deviation from this form of words seems particularly strongly to suggest that the proposition in "all" cannot be true.

In practice, we might argue that it would be odd here for John to exhibit priming effects to the extent of reusing all the lexical material: instead, if the answer to Mary's question in (1) was affirmative, John could just say "yes". The fact that he does not could thus be argued to implicate that the answer must be "no", and therefore that "not all" is the case, much as in the account based on contrastive enrichment specified above[1]. Once again, none of this applies to the case of (2), in which the stronger proposition ("…all…") is not mentioned in the preceding context in any way.

In short, there are several possible explanations for the ostensibly "relevance-based" implicature failure in (2) by comparison with (1). In the following section, I will discuss some recent data concerning implicatures in the numerical domain, and argue that this richer domain represents a better testbed for distinguishing between these competing accounts.

Suppressing irrelevant implicatures from numerical expressions

Cummins, Sauerland and Solt (2012) provide experimental data suggesting that utterances such as (3) convey upper-bound readings to the effect that (4). This suggests that "more than *n*", for numeral *n*, might give rise to a quantity implicature. This would contradict the claim of Fox and Hackl (2006) that "more than *n*" fails to give rise to such implicatures, a claim supported by the observation that (5) does not convey (6).

(3) I have more than 60 CDs.

(4) The speaker does not have more than 80 CDs.

(5) John has more than two children.

(6) John does not have more than three children.

[1] We might ask why John's failure to say "no" directly doesn't implicate that the answer is "yes". There are several possible explanations: one simple possibility is that direct denial is disfavoured on politeness grounds, so its non-occurrence does not have similar pragmatic consequences to the non-occurrence of direct agreement.

Cummins et al. (2012) also document that the utterance (3) does not so robustly convey the upper-bound reading (4) if it occurs in a context in which the number has already been mentioned, for example if (7) was the preceding discourse turn.

(7) This case holds 60 CDs. How many CDs do you have?

They propose that "more than 60" usually gives rise to an implicature conditioned by numeral salience. A speaker who was able to assert "more than 80" would prefer this over "more than 60" on the basis that it is more informative, while also using a number (80) that is equally salient to the alternative (60) and therefore cognitively inexpensive to use (following work on the psychology of number: see for instance Dehaene 1998, Butterworth 1999). By contrast, "more than 61", although more informative than "more than 60", uses a number that is less salient, and by hypothesis more costly to use in processing terms. For this reason, a hearer who encounters "more than 60" can conclude that the speaker is not able to assert "more than 80" but cannot conclude that the speaker is unable to assert "more than 61" – in fact, the speaker may not have done so simply to avoid using this less available number.

This argument can also be expressed in terms of RT. We can assume that the processing of a salient number involves less cognitive effort than that of a non-salient number. Consequently, the use of "more than 61" – even if true – would involve the hearer incurring additional cognitive effort for the gain of relatively little extra cognitive effect (specifically, excluding the possibility that "61" might be the case) . Hence, "more than 61" might not be more relevant than "more than 60". By contrast, the use of "more than 80" would involve no extra cognitive effort and contribute additional cognitive effects, thus achieving higher relevance than "more than 60". It is plausible, therefore, that the hearer of "more than 60" (who under RT's assumptions presumes that this utterance is optimally relevant) can infer that “more than 80” must not be the case, but cannot arrive at any firm conclusion about whether or not “more than 61” is the case.

The weakening of the implicature “60” +> “not more than 80” when the prior context already makes the numeral salient (as in (7)) can be accounted for in at least two separate ways, either of which can be enfolded within an RT explanation. One possibility is that the number 60 is primed by its prior use in the discourse, and consequently can be reused with less effort (both on the part of the speaker and of the hearer). Under these circumstances, “more than 60” may be the most relevant possible utterance even if “more than 80” is the case: what it loses in cognitive effects it gains in the saving of cognitive effort. A rational hearer should, on that basis, be somewhat less willing to conclude that “more than 80” does not hold, given the utterance of “more than 60”, than they would be if the latter was uttered out of the blue.

An alternative explanation is that the failure of the implicature results from QUD considerations. In Cummins et al.’s materials, it could be argued that the prior context invites the hearer to infer the existence of a more specific QUD. In the case of (7), we might infer that the QUD that is raised is whether or not the addressee has more than 60 CDs, rather than merely how many CDs the addressee has. Interpreted against that backdrop, (3) could be considered as a direct answer, in which the speaker affirms that it is the case that they have more than 60 CDs.

From an RT standpoint, we can interpret this idea as a claim about cognitive effects. If the primary discourse goal is simply to answer the QUD posited above, then there is no significant gain to be made in terms of cognitive effects by performing further pragmatic enrichment, and this would come at a cost. Assuming that the role of (3) is to maximise relevance, we should, by this account, take it to convey “more than 60” and not to implicate any kind of upper bound. As discussed earlier, we might consider the QUD as a guide to when to stop reasoning about the utterance: that is, if we know that the QUD is answered, we can immediately cease considering the possibility of pragmatic enrichment, confident that any such enrichment will not yield appreciable gains in cognitive effects (and will incur additional cognitive effort, and thus have lower relevance than the non-enriched interpretation).

Against this, we could argue that QUDs do not arise in isolation: indeed, Roberts (2012) construes discourse as giving rise to a stack of QUDs, such that the ultimate discourse purpose is satisfied by answering all these QUDs. Returning to the exchange (7)/(3), we can naturally construe this as taking place within a dialogue in which one party is trying to suggest a suitable CD case for the other to buy. If this is so, then although "whether more than 60" may be the immediate QUD, the question of "how many" remains unanswered by (3). Concretely, if the customer has 110 CDs, for them to say "more than 60" invites a further dialogue subsequence in which they are offered, say, a case with room for 80 or 100 CDs, and have to explain that this is inappropriate – a subsequence that could be circumvented entirely if they simply stated how many CDs they have at the first opportunity.

A complication here is that Roberts (2012) discusses answers to QUDs in terms of contextual entailment: that is, "a *complete answer* is a proposition which contextually entails an evaluation for each element of q-alt(*q*) [the set of alternatives]" (ibid., 11). Under that definition, (9) and (10) would both be complete answers (in the affirmative) to the (polar) QUD (8).

(8) Do you have more than 60 CDs?

(9) I have more than 60 CDs.

(10) I have more than 300 CDs.

For (10), the proposed heuristic that we should stop enriching once the QUD is answered appears not to be appropriate. The speaker of (10) seems to convey that she has between 300 and 400 CDs, but that enrichment is not relevant to determining the answer to (8). An intuitive way of looking at this is that the speaker of (10) has already elected to provide much more information than is necessary to answer the QUD, and – if cooperative – must therefore be presumed to be answering additional potential QUDs[2]. This might account for why that heuristic would not apply: given that additional QUDs are in play, there is no straightforward way to characterise whether additional pragmatic enrichment will achieve greater relevance. Consequently, applying the heuristic may not result in maximal relevance being achieved. Whether it did so would depend on what the other QUDs were and how well they would be answered by the enriched versus the unenriched utterance – which is a problem we can frame within RT, although the answer is likely to be complex. By contrast, (9) does not signal that the speaker is attempting to answer any QUDs apart from (8), so the decision not to enrich the meaning of (9) pragmatically would appear to be a sensible one.

In short, then, the idea that the satisfaction of a particular QUD results in the suppression of further pragmatic reasoning appears plausible. However, in the case of numerical expressions, the issue then arises of what additional QUDs are in play and how these can be identified. I turn to this topic in the following section.

Inferring contexts from the utterance

[2] This would follow if we assume that over-informing is disfavoured *per se*, as per Grice’s second quantity submaxim. It would also follow if we assume that the hearer of (10) has to perform additional reasoning steps in order to arrive at an answer to (8), whereas the hearer of (9) obtains that answer at once. This latter idea is not entertained by Roberts, who draws no distinction between direct and indirect entailments, but could nevertheless be expressed within RT in terms of (10) imposing additional processing costs on the hearer. Which analysis to adopt depends on whether we construe the speaker of (10) as telling the hearer extra things that they do not obviously need to know, or telling the hearer the things they need to know but in an unhelpfully indirect fashion.

Researchers in experimental semantics and pragmatics have long noted the possibility of their participants inferring the existence of specific contexts, given a particular utterance. Breheny et al. (2006: 445) point out that "even single-sentence utterances can create their own context through a variety of presupposition triggers and information-structure triggers". As a methodological issue, it certainly seems clear that we cannot present dialogue fragments out of context and expect experimental participants to judge them *in vacuo*: and experimental designs (including those of Breheny et al. 2006) have attempted to address this problem. However, the question of precisely how these contexts come to be inferred, and what they consist of, has attracted less attention.

The literature on QUD notably reflects an interest in how the sentence itself, as well as the prior context, contributes to the identification of the QUD. As discussed earlier, the view adopted by Roberts (2012) is one in which the QUD may be inferred by discourse participants even when it is only implicit in the context. Roberts (ibid.: 8) sees this as a special case of the process of plan inference, whereby we "infer interlocutors' plans from other information in the common ground plus what is actually said". That is, the QUD can be identified partly on the basis of what has gone before and partly on the basis of what is happening in the current utterance. This idea has been explored particularly with reference to focus structure, where the observation is that sentential focus is associated with the material that is being questioned (Rooth 1996). We may be able to identify what is being questioned from the prior context, but we may also be able to detect the focus structure of the sentence from the signal itself, and use this as a cue to identifying the QUD. As Zondervan (2011: 225) puts it, "a sentence-level property like focus structure is actually reflecting a contextual property, namely that of the QUD". But, as Roberts (2012) identifies, this is just one of the factors at play. She implies that other sentence-level considerations may also enter into the hearer's calculation of the QUD, and indeed that other inferences about the goals of the discourse may also be drawn on the basis of the current sentence's content.

In the case of numerically quantified expressions, it seems quite clear that more specific QUDs than simply "how many" can be signalled by, and recognised from, the use of specific numbers and quantifiers. For example, (11), (12) and (13) each use the quantifying expression "more than 60" with the argument "senators".

(11) Perhaps by then, more than 60 senators will side with victims over commanders.[3]

(12) More than 60 senators prepared to vote for Rx importation.[4]

(13) "There's more than 60 senators, I'm convinced, who are prepared to vote for this bill, including Don't Ask Don't Tell".[5]

These three examples concern different topics within US politics, but each appears to answer a related QUD: in each case, not merely how many senators are predicted to vote in favour of a particular measure, but specifically whether or not more than 60 are predicted to do so. For this reason, each of (11)-(13), even taken out of context, seems strongly to suggest that this particular question is of broader interest – that is, that 60 is a critical threshold of votes in the US Senate for these various purposes (as is in fact the case).

It also seems natural to interpret these examples as not conveying a clear pragmatic upper bound, which would be a logical consequence of the QUD being "whether more than 60", as discussed in the preceding section. This seems to agree with the results obtained by Cummins et al. (2012): they documented somewhat inconsistent behaviour with "more than

[3] http://www.mysanantonio.com/opinion/editorials/article/Setback-for-sexual-assault-victims-5297981.php, retrieved 27 May 2014

[4] http://www.fdanews.com/articles/71127-more-than-60-senators-prepared-to-vote-for-rx-importation-enzi-says, retrieved 27 May 2014

[5] http://www.nydailynews.com/news/politics/don-don-joe-lieberman-democrats-60-votes-repeal-policy-article-1.452082, retrieved 27 May 2014

n" when there was no numeral mentioned in the prior context, and particularly when the number *n* itself was non-round (for instance, 97). That could be attributable to some of the participants inferring contexts of utterance in which "more than *n*" was in fact a QUD, and declining to perform pragmatic enrichment for that reason. Indeed, in corpora such as the BNC, the use of "more than *n*" with large non-round *n* is extremely rare in cardinal contexts, and appears to be largely restricted to cases in which the specific numeral is of especial importance in the given context.

What about the use of small numbers in conjunction with "more than", as in (5), repeated below? It seems uncontroversial that (5) does not implicate (6) – we can discern this on the basis that (5) and (6) taken together would entail (14), and it is uncontroversial that an utterance of (5) does not typically convey (14).

(5) John has more than two children.

(6) John does not have more than three children.

(14) John has exactly three children.

There are, in principle, several reasons why this might be the case. First, we could argue that "more than two" systematically fails to give rise to implicatures of this kind, perhaps on the basis that small numbers are mentally represented differently to large numbers in some way. However, this seems untenable: the prohibition on implicature only seems to extend to cardinal usages, whereas (15) does seem to implicate (16). It would seem odd to require different analyses for the integer and real-valued cases.

(15) On average, families in Europe have more than two children.

(16) It is not the case that, on average, families in Europe have more than three children.

A second possibility is that the candidate implicature from (5) fails on the basis of its inherent inefficiency. We could reason as follows: if (5) implicated (6), the use of (5) by an informed and cooperative speaker would convey (14). However, no such speaker would choose to convey (14) by uttering (5), as this achieves no greater effect and requires more effort on the part of the hearer (alternatively, because (5) is informationally weaker and no easier to express). As the hearer, we should rationally conclude that the speaker cannot possibly have meant that, and dispense with the implicature for that reason.

This line of analysis resembles the multi-step process of reasoning proposed by Degen, Franke and Jäger (2013) for the assignment of referring expressions. It certainly seems computationally tractable for us, as hearers, to proceed in this way. Nevertheless, as far as example (5) is concerned, such an account is strongly counterintuitive. When hearing (5), we appear to latch on automatically to the idea that the utterance conveys only a lower bound reading, and there is no evidence that we need to go through a complex reasoning process to cancel the unwanted implicature.

A third and more intuitive possibility is that the use of "more than two" in a cardinal context conveys that the specific QUD is "whether or not more than two is the case". If we assume that the speaker of (5) is knowledgeable and cooperative, it appears at least very likely that this kind of QUD is in effect: if the actual QUD existed at a higher or lower level of specificity, then the use of (5) by a knowledgeable speaker would supply either too much or too little information. So the hearer of (5) might reasonably conclude that either the QUD concerns whether or not "more than two" is the case, *or* that the speaker is not fully knowledgeable on the topic, *or* that the speaker is uncooperative. In any of these cases, the rational hearer should decline to perform a pragmatic enrichment.

This account squares with the observation that the speaker of (17) would naturally be assumed to be in a context in which the question of whether or not John has more than two children is sufficiently salient to license it being answered directly.

(17) John has more than two children; in fact, he has five.

However, if we are to be able to account for examples such as this, we may need to refine the notion of what constitutes an answer to a QUD. For (17), even if the current QUD concerns whether or not John has more than two children, simply affirming that he has five children would clearly entail a complete answer to this question. The justification for the circumlocution of (17) seems to be that it is helpful to give a direct answer to the QUD, and not to require the hearer to perform an additional inference (figuring out that "five" is "more than two"). However, this notion of a "direct answer" being preferable seems to conflict with the idea that contextual entailment is a sufficient condition for complete answerhood.

Distinguishing QUD from priming effects

There is at least one more distinct account of how the implicature from "more than two" might systematically fail to arise. On this account, rather than inferring the existence of a specific QUD of the kind sketched above, the hearer of (5) simply infers that the speaker had a particular reason to use this precise number (or, more generally, expression). Consider the use of "more than 100". We might argue that this doesn't typically reflect the existence of a QUD "whether more than 100", but simply represents a cognitively efficient way to convey an approximate value, given the high salience of "100" in our mental number system compared to its near neighbours. Analogously, if we imagine circumstances in which "two" was made much more salient than its neighbours, we might expect a speaker to use it in preference to its neighbours, even if the question of "whether more than two" was not pertinent to any of the participants' discourse goals.

Under such circumstances, the hearer would be predicted not to draw pragmatic enrichments, on the basis that the speaker may have declined to utter informationally stronger alternatives merely on the basis that they did not involve primed numerals. Assuming that a primed numeral is, by hypothesis, easier to use than an unprimed numeral, the use of informationally stronger alternatives would then have come with an additional cost (potentially both for speaker and hearer), which might not have been outweighed by the additional benefits in terms of cognitive effects. That is, it is possible that the most relevant option (in an RT sense) would be for the speaker to make the weaker statement using the primed numeral. Hence, the rational hearer of "more than two" should not attempt to perform pragmatic enrichment in this case.

I take it as uncontroversial that this conjectured priming effect is not the same thing as adherence to QUD, even though the pragmatic consequences would be superficially similar. Priming is taken to be a low-level, unconscious process, whereas the establishment of the QUD is a matter of high-level discourse planning. If the speaker is assumed to be providing a minimal and direct answer to the QUD, the hearer can further infer something about the interlocutors' discourse goals, whereas if the speaker is assumed merely to be exhibiting priming effects, the hearer can infer something about the content of the previous discourse, but nothing about its goals. Priming effects of this kind could, in principle, arise purely by happenstance: if a number were to be mentioned frequently in the context of one set of discourse goals, it could in principle exert an effect on the numbers used in a separate portion of the discourse, even if it were not relevant to any of the goals of this latter discourse.

If the intuitions discussed earlier in this paper are correct, the ability of hearers to infer the existence of a specific QUD are not in doubt. The question of whether they hearers also take account of the possible presence of priming effects remains open. Rationally, this should be bound up with the question of whether speakers exhibit priming effects of this kind in their productions. Although the rationale for priming effects in this domain appears sound, it may be extremely marginal in actual discourse. Notably, most successful demonstrations of priming effects appear to have focused on areas with relatively low informational load: for instance, dative alternation (Branigan et al. 2000). Although speakers may have a clear preference in this domain (e.g. for saying "Mary gave the X the Y" rather than "Mary gave the Y to the X"), and this preference may be subordinated to priming of the disfavoured form, the choice of expression does not seem to have any discernible communicative consequences. By contrast, saying "more than two" rather than "more than three" would involve giving up a substantial amount of information, and it is not clear whether priming effects are strong enough to compel speakers ever to do so. Intermediate cases are those in which there is a relatively free choice of number, and the possibility of priming effects might account for examples such as (18)-(20). (Note that "24 carat" denotes a purity standard for gold, but merely a weight for diamonds, which carries no especial significance.)

(18) He's a 24-carat diamond geezer[6]

(19) How much does a 24 carat diamond cost?[7]

(20) Damien James – 24 Carat Diamond[8]

To distinguish between the QUD and priming-based explanations for implicature failure in cases such as (5) would require further experimental work, which I am not attempting within this paper. I conjecture that it will not be possible to disprove the priming

[6] http://www.theguardian.com/artanddesign/2007/jun/10/art, retrieved 27 May 2014
[7] http://wiki.answers.com/Q/How_much_does_a_24_carat_diamond_cost?, retrieved 27 May 2014
[8] http://www.trevorgeorge.co.uk/portfolio/damien-james-24-carat-diamond/, retrieved 27 May 2014

account directly via experimental means, for the following reason. Suppose that the QUD-based account is correct and the hearer of (5) assumes that the question of “whether more than two” is open. We can attempt to explore priming effects by manipulating the prior context to include relevant or irrelevant instances of the number “two”. However, given that the speaker is knowledgeable and cooperative, I would argue that (5) so strongly suggests the presence of an open QUD “whether more than two” that the contextual manipulation will have no effect.

We can nevertheless go some way towards testing these competing accounts by using materials with larger round numbers. Recall that we expect “more than 60” to implicate “not more than 80”, but for this effect to be vulnerable to contextual manipulation. That is, when “60” is mentioned in the prior context in a relevant way, the implicature is less reliably available to hearers (as shown by Cummins et al. 2012). If the priming explanation is the correct one, this effect should if the numeral “60” is merely mentioned in the prior context, even in a way that is irrelevant to the quantifying statement under test. If the QUD-based explanation is correct, the effect should be available only if the numeral “60” is used in such a way as to introduce the question “whether more than 60” into the discourse. Hence, by manipulating the status of the numeral in the preceding context, we can go some way towards discerning whether the reduction in implicature is due to priming or to QUD-driven effects. In either case, participants may still infer the relevance of “60” or the existence of the QUD “whether more than 60”, and suppress implicature on that basis, so the results might not be clear-cut. However, as we know that “more than 60” can implicate “not more than 80” when presented out of the blue, we can be confident that the suppression of implicature is not obligatory given this quantifying expression.

Summary and conclusion

There is ample evidence that quantity implicatures are suppressed in cases where the stronger statement, which the implicature negates, would not have been relevant. In principle, this suggest that pragmatics can play an important role in elucidating the higher-level structure of dialogue. However, it is correspondingly difficult to define precisely what constitutes "relevance". Relevance Theory offers an account in which optimal relevance is construed as the greatest possible ratio of cognitive effects to effort on the part of the hearer. Yet even if we adopt this construal of relevance, there are at least two distinct reasons why an implicature might not be obtained by the hearer in a given situation. It could be that the QUD is already satisfied without further enrichment, and thus obtaining the implicature would yield insufficient cognitive effects to be justify the additional effort. Or it could be that the use of a weak expression is more economical in terms of cognitive effort than the stronger alternative would be, under the particular circumstances that prevail at the time of utterance.

As is so often the case in experimental studies of implicature, the relevant literature has focused on very few triggers, predominantly "some" (+> "not all") and "or" (+> "not and"). A typical approach has been to contrast the interpretation of an implicature trigger in a context in which the stronger alternative is explicitly made relevant with its interpretation in a more neutral context. However, this approach does not enable us clearly to distinguish the two accounts sketched in the previous paragraph: indeed, some of the materials admit a third interpretation in which the pragmatic enrichment is potentially attributable to contrast effects. Moreover, when the triggers are presented in ostensibly neutral contexts, it is likely that hearers sometimes infer the existence of more elaborate contexts and interpret the target expression as though it had been produced in that specific context.

In this paper I have discussed the possibility of using numerical expressions as a testbed for investigating which aspects of the prior context are materially relevant to the question of whether or not an implicature is recovered. Expressions such as "more than *n*", used in cardinal quantifying contexts, give rise to implicatures, and these implicatures appear to be susceptible to contextual manipulation in the predictable way. In some respects the pragmatics of such utterances is more complex than that of implicature triggers like “some”: various distinct enrichments are available, an assortment of different contexts can be inferred by hearers, and these issues are bound up with the psychology of number in a non-trivial way. However, this does present a potentially rich domain of experimental enquiry. Notably, we can manipulate the role as well as the presence of numerals in the prior context, and thus distinguish between the effect of providing an explicit QUD involving a particular numeral and the effect of merely mentioning that numeral. We can also exploit the observation that different numerical expressions give rise to different inferences about the content of the prior discourse, and use this to explore how we integrate inferences of this kind with the explicit knowledge that we have about the discourse. This has the potential to contribute to a fuller understanding of how we infer QUDs (as well as other aspects of the overarching discourse purpose), and how we update those inferences in response to new incoming information.

Research on low-level priming effects, striving towards a mechanistic psychology of dialogue (as expressed by the title of Pickering and Garrod’s 2004 paper), posits that speakers’ productions are substantially conditioned by the content of the preceding discourse. Accounts based on high-level planning, such as Roberts’s (2012) model, take a view of communication that is markedly more strategic on the part of the interlocutors. By better understanding the respective roles of QUD-based and priming-based effects in the selection and interpretation of expressions, we can obtain a fuller appreciation of the relative contributions of these two features of dialogue to the process of human-human interaction.

References

Bonnefon, Jean-François, Aidan Feeney, and Gaelle Villejoubert (2009). When some is actually all: scalar inferences in face-threatening contexts. *Cognition, 112*: 249–58.

Branigan, Holly P., Martin J. Pickering, and Alexandra A. Cleland (2000). Syntactic coordination in dialogue. *Cognition, 75*: B13–25.

Breheny, Richard, Napoleon Katsos, and John N. Williams (2006). Are scalar implicatures generated by default? *Cognition*, *100*: 434–463.

Butterworth, Brian (1999). *The Mathematical Brain*. London: Macmillan.

Carlson, Lauri (1982). *Dialogue Games: An Approach to Discourse Analysis*. Dordrecht: D. Reidel.

Chemla, Emmanuel, and Benjamin Spector (2011). Experimental evidence for embedded scalar implicatures. *Journal of Semantics, 28*(3): 359–400.

Chierchia, Gennaro (2006). Broaden your views: implications of domain widening and the “logicality” of language. *Linguistic Inquiry*, *37*: 535–590.

Cummins, Chris, and Napoleon Katsos (2010). Comparative and superlative quantifiers: pragmatic effects of comparison type. *Journal of Semantics, 27*: 271–305.

Cummins, Chris, Uli Sauerland, and Stephanie Solt (2012). Granularity and scalar implicature in numerical expressions. *Linguistics and Philosophy, 35*: 135–69.

Davies, Catherine, and Napoleon Katsos (2010). Over-informative children: production/comprehension asymmetry or tolerance to pragmatic violations? *Lingua*, *120* (Special issue on Asymmetries in Child Language): 1956–72.

Degen, Judith, Michael Franke, and Gerhard Jäger (2013). Cost-based pragmatic inference about referential expressions. In Markus Knauff, Michael Pauen, Natalie Sebanz and Ipke Wachsmuth (eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 376–381). Austin, TX: Cognitive Science Society.

Dehaene, Stanislas (1997). *The Number Sense*. New York: Oxford University Press.

Fox, Danny and Martin Hackl (2006). The universal density of measurement. *Linguistics and Philosophy*, 29: 537–586.

Geurts, Bart, and Nausicaa Pouscoulous (2009). Embedded implicatures?!? *Semantics & Pragmatics*, *2*(4): 1–34.

Geurts, Bart, and Bob van Tiel (2013). Embedded scalars. *Semantics & Pragmatics*, *6*(9): 1–37.

Goodman, Noah D., and Andreas Stuhlmüller (2013). Knowledge and implicature: modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*: 173–84.

Grice, H. Paul (1975). Logic and conversation. In Peter Cole and Jerry L. Morgan (eds.), *Syntax and Semantics*, Vol. 3 (pp. 41–58). New York: Academic Press.

Leech 1983

Levinson, Stephen C. (2000). *Presumptive Meanings*. Cambridge, MA: MIT Press.

Pickering, Martin J., and Simon Garrod (2004). Towards a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*: 169–226.

Roberts, Craige (2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics & Pragmatics, 5*(6): 1–69.

Rooth, Mats (1996). Focus. In Shalom Lappin (ed.), *Handbook of Contemporary Semantic Theory* (pp. 271-297). Oxford: Blackwell.

Van der Henst, Jean-Baptiste, Laure Carles, and Dan Sperber (2002). Truthfulness and relevance in telling the time. *Mind and Language*, *17*: 457–466.

Zondervan, Arjen (2011). The role of QUD and focus on the scalar implicature of *most*. In Jörg Meibauer and Markus Steinbach (eds.), *Experimental Pragmatics/Semantics*, (pp. 221–238). Amsterdam: John Benjamins.