A Speaker-Referring OT Pragmatics of Quantity Expressions

Chris Cummins

SFB 673 – Alignment in Communication, Universität Bielefeld, Germany

c.r.cummins@gmail.com

Abstract. Constraint-based approaches to pragmatics have customarily focused on the hearer, and offered accounts of the optimal interpretation of utterances. Blutner (2006, i.a.) has argued that it is necessary also to consider the role of the speaker, and thus motivates a bidirectional Optimality Theory (OT) approach to pragmatics. However, as he notes, this may have limited explanatory potential from a processing viewpoint. A recent account, focusing on expressions of quantity, aims instead to model the speaker's choice of expression by means of unidirectional OT mechanisms. In this paper I discuss the merits of this approach versus OT-based alternatives. In particular, I explore the implications of this account for the hearer, who is required to solve the problem of utterance interpretation in a non-OT fashion in this model, and explore how this task can be made tractable. I briefly discuss the novel predictions about interpretation and processing that arise from such a development, and consider the theoretical implications for pragmatics in general.

Keywords: Pragmatics, quantity expressions, implicature, OT, granularity

1 Introduction

One of the most general problems in linguistic pragmatics is how enrichments to the meaning of an utterance are computed by hearers, given that – at least in principle – virtually any utterance may be used to convey virtually any meaning. For example, in (1), B's response can naturally be interpreted as answering A's question in the negative, even though B does not use any negative form or make any reference to the content of A's utterance.

A: Is Bill a good person to work for?
B: Has anyone shown you where the print room is? (1)

Within classical pragmatic approaches, this interpretation is considered to be cued by the apparent flouting of a conversational maxim which would require that B's contribution was relevant to "the accepted purpose or direction of the talk exchange" (Grice 1975: 45). Given the apparent lack of relevance of B's utterance, A is entitled to undertake reparatory inferences that enrich its meaning. In this particular case, A

© Springer-Verlag Berlin Heidelberg 2011

might reason as follows: given A's polar question, it would be cooperative for B to respond "yes" or "no" (or "I don't know"). These options would also be economical in terms of the effort involved for both speaker and hearer. B's failure to use such a form suggests that B is unwilling to commit to any of these statements. As they cannot all simultaneously be false, at least one of them must be blocked for some other reason. Noting that a "no" response would be impolite, a plausible abductive inference is that B considers "no" to be the correct response to A's question.

If we accept this kind of account as broadly satisfactory, it is nevertheless unclear as to how the reasoning processes involved can be modeled and computationally implemented. Such difficulties are shared by more recent philosophical pragmatic accounts such as Relevance Theory (Sperber & Wilson 1986/1995). Within this paradigm, hearers are presumed to perform pragmatic enrichments in such a way as to achieve optimal relevance, where this is defined as maximizing the ratio of cognitive effects to effort. Although this descriptively appears to match our intuitions about many kinds of pragmatic inference, it does not specify how hearers can practically go about achieving this result.

One recent approach to the problem of characterizing pragmatic enrichments of this type is to appeal to the notion of constraint satisfaction. In this paper, I focus on an approach that considers the speaker's choice of utterance to reflect the satisfaction of multiple soft constraints. In particular, I explore the implications of this for the hearer, and consider both the inferences that are available to the hearer in principle and how they might be recovered in practice.

The rest of the paper is organized as follows. In section 2, I briefly survey constraint-based approaches to pragmatics, and argue for the usefulness of speaker-referring constraints. In section 3, I outline a recent unidirectional speaker-referring account of the production of numerical quantity expressions. In section 4, I consider the nature of the pragmatic enrichments that are predicted to be available to the hearer on this account. In section 5, I explore some possible mechanisms that the hearer might use in order to achieve the potential pragmatic enrichments, and sketch the implications of this for pragmatic theory. Section 6 provides a brief summary of the paper.

2 Constraint-based approaches to pragmatics

2.1 The hearer-referring approach

One way of using constraints in pragmatics is to posit that the interpretation of a given utterance reflects the hearer's attempt to satisfy multiple competing soft constraints. Hendriks and de Hoop (2001) treat anaphora resolution in this way, focusing on how the hearer of the expression "who wants one?" establishes the intended referent of "one". They argue that the process involves dealing with a complex of contextual, syntactic and intonational constraints, whose interaction can be modeled within Optimality Theory. Their approach is a unidirectional hearer-referring one, which takes a linguistic utterance as its input and delivers an interpretation as its output. From this perspective, we are not interested in the authorship of the utterance: it is

taken as given, and the model attempts to characterize the process by which the hearer retrieves the meaning of the utterance.

2.2 The bidirectional approach

Contrary to the assumptions of the hearer-referring approach, Blutner (2000) argues, on the basis of examples of blocking and pronoun binding, that it is generally necessary also to take the speaker's needs into account. In this way, he motivates a bidirectional OT mechanism. Crucially, bidirectional OT has an additional level of explanatory power. In bidirectional OT, it is possible to explain the inadmissibility of a particular interpretation on the basis that, if the speaker had intended to convey that meaning, they would have encoded it in a different way. From a pragmatic standpoint, this kind of bidirectionality is particularly relevant to the treatment of markedness implicatures, discussed by Horn (1984), Levinson (2000) and others. It has long been argued, for instance, that an utterance such as (2) is interpreted as indicating that the action of "causing to die" was performed in an indirect way, on the basis that otherwise it would have been expressed by the word "kill".

The outlaw caused the sheriff to die. (2)

However, the processing implications of a bidirectional account are not straightforward, as Blutner acknowledges. There are two ways of obtaining markedness implicatures within a bidirectional OT pragmatics. The first involves positing linking constraints which explicitly stipulate that marked forms should convey marked meanings, an approach which is unattractively *ad hoc* (Blutner 2000: 10). The second involves appeal to the notion of "weak bidirection", which is more permissive than the strong form. In the strongly bidirectional model, a form-meaning pair <f, m> is optimal if there exists no alternative meaning m' and no alternative form f' such that <f, m'> or <f', m> better satisfies the constraints than <f, m>. In the weakly bidirectional model, a form-meaning pair <f, m> is "super-optimal" and selected if there exists no alternative meaning m' or form f' such that either <f, m'> or <f', m> **both** satisfies the constraints better than <f, m> **and** is itself super-optimal.

To unpack this formalism for example (2): let $KILL_1$ denote killing in a direct way, and $KILL_2$ denote killing in an indirect way. We want our model to select the form-meaning pair <cause to die, $KILL_2$>. For this to be optimal, in a strongly bidirectional sense, it must better satisfy the ranked constraints of the model than either <cause to die, $KILL_1$> or <kill, $KILL_2$> manage to. However, assuming that the form "kill" is less marked than the form "cause to die", <kill, $KILL_2$> is more satisfactory than <cause to die, $KILL_2$> with respect to markedness constraints. Similarly, assuming that the meaning $KILL_1$ is less marked than the meaning $KILL_2$, <cause to die, $KILL_1$> is more satisfactory than <cause to die, $KILL_2$> with respect to markedness constraints. The only way that <cause to die, $KILL_2$> can win the competition against both these alternatives, and thus be optimal, is if the competing pairings are both blocked by some linking constraint: for instance, one prohibiting the association of marked forms with unmarked meanings and vice versa. However, positing such a

constraint amounts to stipulating within the model that markedness implicatures occur, and consequently lacks explanatory power.

Contrastingly, for the form-meaning pair <cause to die, $KILL_2$> to be super-optimal (as required in the weakly bidirectional model), all that is required is for neither of the competitors <cause to die, $KILL_1$> and <kill, $KILL_2$> to be super-optimal. Both of these pairs are in competition with <kill, $KILL_1$>, which is itself super-optimal; therefore neither of the pairs are super-optimal, and it follows that <cause to die, $KILL_2$> is super-optimal, as required. From this perspective, we can thus explain the association of the marked form with the marked meaning without appeal to linking constraints.

Such an account remains problematic on several levels. For convenience, we have assumed in this demonstration that there is only one marked form and one marked meaning to be paired, and the generalization of the approach to larger systems may not be straightforward. More pertinently for the purposes of this paper, it is acknowledged by Blutner that the system of weak bidirection is not appropriate for systems that are intended to be cognitively plausible, on the basis that it requires global solution. Hence, it "[does] not even fit the simplest requirements of a psychologically realistic model of online, incremental interpretation" (Blutner 2006: 16). A bidirectional account that is to be psychologically plausible must, it appears, adopt a system of strong bidirection at the cost of positing highly stipulative linking constraints that vitiate the account's explanatory power.

2.3 The speaker-referring approach

A third logical possibility, pursued less often in prior work, is to consider the role of the speaker as primary. If we assume that the hearer's task is to attempt to reconstruct or decode the speaker's communicative intention, then authorship of the utterance assumes particular importance: the speaker determines which meanings are present and should be deciphered. For instance, in the "who wants one?" example, the meaning of "one" appears uncontroversially to be "whatever the speaker meant by it" – the speaker has a particular referent in mind at the time of encoding, and this is not negotiable.

From this perspective, the communicative success of a hearer-referring OT system depends upon the speaker's intention being encoded in a way that is decipherable by this system, but no account is offered as to how this non-trivial task can be accomplished. Instead, this approach considers the utterance to be 'given', and implicitly assumes that it is related to the speaker's intention in some appropriate way. However, the utterance does not in fact arise *in vacuo* and its construction is at least chronologically prior to its interpretation. It could be more effective to posit instead that the speaker is engaging in a constraint-governed process of production and it is the hearer that performs the complementary role of 'undoing' this process to reconstruct the intended meaning. Such an approach resembles the Dual Optimization of Smolensky (1996), in which production and comprehension are distinct processes, more closely than it does the simultaneous optimization of bidirectional OT.

In the following section I discuss a recent proposal focused on the pragmatics of numerically-quantified expressions, which adopts the speaker-referring unidirectional stance. I consider its relation to bidirectional accounts of numerical quantification. In particular, I examine the obligation that it places on the hearer to unpack the various aspects of communicated meaning, and critically assess how that obligation could be met. I will argue that this model has some explanatory advantages with respect to recent data, and that it gives rise to potential testable hypotheses about the nature of the hearer's reasoning process.

3 An OT production model of numerical quantity expressions

Cummins (2011) proposes a speaker-referring unidirectional OT model of the production of numerically-quantified expressions. This domain has been studied from a constraint-based perspective, and specifically by appeal to bidirectional OT (Krifka 2002, 2009). It is particularly amenable to a constraint-based treatment for several reasons. First, there are frequently numerous semantically truthful candidate expressions corresponding to a particular communicative intention, because of the rich entailment relations that exist among numerical expressions. Under any circumstance in which "more than 100" can truthfully be uttered (within a normal quantificational context), it would also be true to say "more than 99", "more than 98", and so on. As such descriptions are not always maximally informative, there is a non-trivial mapping problem to address here. Secondly, the numerical domain offers a convenient means for quantifying certain types of constraint violation. For instance, we could think of "more than 99" as being one unit less informative than "more than 100", in that (in the cardinal case) it admits precisely one more possibility for the value of the quantity under discussion: in the former case, but not the latter, the quantity might be exactly 100. A third consideration is that the use of numerical quantifiers has been argued to depend upon a wide range of considerations, from aspects of philosophical semantics and linguistic pragmatics through to numerical psychology. These considerations have traditionally been addressed within distinct fields of enquiry, with relatively little cross-talk, despite occasional appeals for interdisciplinary work. A constraint-based account offers a convenient means to attempt to capture all these competing factors within a single coherent model.

Drawing upon diverse sources of theoretical and experimental research, Cummins (2011) proposes six constraints on the use of numerically-quantified expressions: informativeness, granularity, numerical salience, quantifier simplicity, numeral priming, and quantifier priming. Two of these, numerical salience and quantifier simplicity, are treated as markedness constraints, as they govern the form of the utterance irrespective of context. The others are treated as faithfulness constraints, as they govern the relationship between the 'situation' (broadly construed) and the utterance. Specifically, numeral and quantifier priming require the reuse of contextually activated material, granularity requires the appropriate level of specificity in the choice of numeral (see Krifka 2009, among others, for a detailed discussion of this property), and informativeness requires the use of a maximally informative expression given the

communicative intention of the speaker. In Cummins (2011) this last constraint is couched in terms of the speaker maximally excluding possibilities that he or she knows to be false.

The application of these constraints in concert provides a source of novel predictions, as well as recapitulating existing predictions under less stipulative assumptions. For instance, this approach offers an alternative pragmatically-based account of the differences between "more than" and "at least" (cf. Geurts and Nouwen 2007, Büring 2008, Cummins and Katsos 2010), which can also be extended to related classes of expressions (such as those discussed by Nouwen 2010). It also sketches an alternative explanation for the preferences for the approximate interpretation of round number words observed by Krifka (2009). Moreover, it enables us to draw novel predictions about the range of usage and interpretation associated with modified round numerals ("more than 100", etc.) which are borne out experimentally (Cummins, Sauerland and Solt 2012).

4 Interpreting the output of the production model

The predictions outlined in the preceding section make reference both to the choice of expression and to its interpretation. Predictions about the former drop out immediately from the speaker-referring model. However, to draw predictions about interpretation we also have to consider the role of the hearer. In the following subsections I discuss how the hearer's role differs within this account to that posited in bidirectional and hearer-referring OT approaches, before proceeding to consider the evidence that hearers do indeed behave in the way that the speaker-referring account most naturally predicts.

4.1 The hearer's task in a speaker-referring model

Under the assumptions of the model discussed above, the speaker's choice of expression depends on various facets of the situation: specifically, those referred to by the constraints. These include the speaker's communicative intention, which is relevant to the evaluation of violations of the informativeness constraint (INFO). However, it also includes factors pertaining to the discourse context, namely which entities are activated or salient within it (in the case of numeral priming, NPRI, and quantifier priming, QPRI), and the level of specificity required (in the case of the granularity constraint, GRAN). Across contexts, the choice of utterance is also argued to depend on the structure of the number and quantifier systems, as modulated by the constraints on numeral salience (NSAL) and quantifier simplicity (QSIMP).

To the speaker, all these considerations can be treated as part of the input to the OT-based decision process, and the linguistic expression produced is the output. The hearer's task, however, is not simply to reverse this process and, upon hearing the output, attempt to figure out the nature of the input (that is, to reconstruct the situation). Instead, the hearer typically shares some or all of the speaker's knowledge about the context, such as which discourse entities are activated, as well as the general

information about the numeral and quantifier systems, such as which numerals are generally salient. The hearer's task is just to infer the information to which s/he is not already privy, the crucial aspect of which is the speaker's intention.

In this way, the hearer's task within this model of this type is unlike their task within any other type of OT model. In a bidirectional model, the system specifies (or permits the calculation of) optimal form-meaning pairings, and the hearer should be able to identify these in exactly the same way as the speaker does. Thus, the hearer can simply read off the optimal interpretation for a given form just as the speaker can read off the optimal form for a given interpretation. In a hearer-referring unidirectional model, the hearer computes a preferred interpretation given the form and the hearer's own interpretative preferences, about which s/he may be presumed entirely knowledgeable. In that scenario, the preferences of the speaker are entirely irrelevant from the hearer's perspective.

4.2 Distinctive interpretation predictions of the speaker-referring model

As a consequence of the above-mentioned differences, the unidirectional speaker-referring account can address aspects of pragmatics that cannot easily be treated within the other formalisms. Consider the case of "more than *n*" for some round *n*, and its preferred interpretations. If the speaker wishes to describe a quantity in excess of 73, say, then (3)-(5) are candidate expressions on the grounds of their semantic truthfulness. For expository convenience, I shall ignore the many other logical possibilities here, and will also assume throughout that semantic considerations automatically rule out false descriptions.

…more than 73… (3)

…more than 70… (4)

…more than 60… (5)

Assuming for the moment that there are no prior contextual commitments, the relevant constraints here are INFO and NSAL. Further assuming, in accordance with the literature on mathematical cognition, that 60 and 70 are more salient than 73, it follows that (3) incurs an additional violation of NSAL. Meanwhile, (4) violates INFO compared to (3), and (5) does so to an even greater extent. Schematically, the resulting tableau is as shown in Table 1 (where the number of asterisks indicates relative, rather than absolute, numbers of constraint violations).

Form	INFO	NSAL
More than 73		*
More than 70	*	
More than 60	**	

Table 1. Tableau for examples (3) to (5), situation "more than 73".

The selected output here depends on the speaker's constraint ranking. With INFO ranked above NSAL, "more than 73" would be preferred; with NSAL ranked above INFO, "more than 70" would be preferred.

Now, by contrast, if the numeral "60" is activated in the preceding discourse, NPRI also becomes relevant, and is violated by those forms which do not (re)use this number. The resulting tableau is given as Table 2.

Form	INFO	NSAL	NPRI
More than 73		*	*
More than 70	*		*
More than 60	**		

Table 2. Tableau for examples (3) to (5), situation "more than 73", "60" primed.

Here, if INFO is ranked above both NSAL and NPRI, "more than 73" is again preferred. However, if NPRI is ranked above INFO, "more than 60" is preferred. Only if NSAL > INFO > NPRI is "more than 70" still optimal.

Taking the hearer's perspective, let us consider the rational interpretation that could be placed on the utterance that is selected under either of these conditions. Crucially, within this model, the hearer may assume that the utterance is optimal given the situation and the speaker's constraint ranking (and given the general assumption that the speaker is behaving cooperatively). Here, 'optimal' specifically means corresponding to the selected output of the speaker's OT system.

Ignoring contextual factors, the hearer might rationally interpret the possible utterances as follows.

- "More than 73" would be optimal only if INFO was the top-ranked constraint, and would therefore reliably signal that the speaker's knowledge extends to the quantity under discussion being "greater than 73". For any other knowledge state, e.g. "more than 74", INFO being the highest ranked constraint would mandate a maximally informative choice of expression.
- "More than 70" would be optimal if NSAL > INFO and the speaker's knowledge was "more than 70", "more than 71", or similarly, up to "more than 79"; or if INFO > NSAL and the speaker knew only that "more than 70" held and could not commit to a higher number. In either case, it would signal that the speaker could not commit to "more than 80", as otherwise "more than 80" would be a preferable output.
- Similarly, "more than 60" could arise only if the speaker was unable to commit to "more than 70".

Note that the model allows a many-to-one mapping between the speaker's intention and the optimal utterance. Under these conditions, it is not possible for the hearer to recapture the precise details of that intention. This again contrasts with the bidirectional model, which seeks optimal form-meaning pairings, and in this case would aim to pair intentions with their preferred means of expression. The unidirectional model can therefore generalize to cases in which distinct intentions are associated with identical preferred forms. It is compatible with the theoretical idea that the speaker may have any of arbitrarily many distinct intentions: for instance, if what the speaker wishes to convey is construed as a probability distribution over possible values. Such a view would seem to permit greater nuance in the speaker's knowledge representation – or at any rate a more transparent mapping between their knowledge and their communicative intention – than could be tolerated within a system in which speaker intentions must be mapped one-to-one to linguistic forms.

Returning to the above example, if we also consider contextual factors, the hearer's task becomes more complex. Suppose first that the hearer knows "60" to be contextually activated (e.g. because they have just asked whether the quantity under discussion is above 60). In this case, the interpretation of "more than 73" or "more than 70" proceeds as above. However, if "more than 60" is uttered, this might reflect adherence to NPRI on the part of the speaker, in which case it is entirely compatible with an intention of "more than 70" or higher values, whereas before it was incompatible with an intention of "more than 70". So in this case, the hearer should not be able to draw the inference that the utterance of "more than 60" signals the speaker's inability to commit to the assertion of "more than 70".

The above discussion presumes that the hearer is completely knowledgeable about whether or not a numeral has been primed in the preceding context. However, it is also possible that a numeral might be primed without the hearer being aware of it. Therefore, any utterance should be interpreted (according to this model) as a potential case of priming. Generally speaking, this will require the hearer to be cautious in drawing conclusions of the type discussed above. The utterance of "more than 73" might in fact reflect the action of a highly-ranked priming constraint in a situation in which the speaker has more precise information but considers 73 to be a significant threshold and worthy of mention. The same applies to "more than 70". Thus the inference arising from these utterances that the speaker cannot commit to "more than 80" should not be exceptionless. The hearer should, however, be able to infer that **either** the speaker is not willing to commit to "more than 80" **or** the numeral used has especial significance (or both). Hence, if the hearer considers it very unlikely that the numeral has especial significance, they should consider it likely that its use conveys something pragmatic about the intended range of interpretation.

4.3 Evidence for constraint-modulated interpretation

It might appear that the model outlined above places an unreasonable burden on the hearer, notwithstanding the convergent evidence that hearers are able to perform complex and highly conditional pragmatic enrichments (cf. Bonnefon, Feeney and Villejoubert 2009; Breheny, Ferguson and Katsos 2013). However, recent experi-

mental data suggests that hearers not only *can* but actually *do* interpret expressions of quantity in this way, at least under certain conditions. Cummins, Sauerland and Solt (2012) demonstrated that the preferred interpretation of an expression such as "more than 70" is indeed one in which a pragmatic upper bound is posited: i.e. the hearer infers that the speaker is unwilling to commit to the truth of "more than 80". Some participants in their study explicitly reported engaging in reasoning of this type. The research further showed that when the numeral 70 is mentioned as a threshold value in the preceding conversational turn, the implicature is attenuated, and the hearer considers that the expression "more than 70" may convey a higher value. This appears to reflect an understanding that, in such cases, the speaker may know that "more than 80" holds but then deliberately choose to make a weaker statement.

In both conditions, the pragmatic upper bounds are inferred frequently but not invariably. This is coherent with the prediction that hearers should factor in the possibility that the use of a specific numeral is deliberate but motivated by considerations to which the speaker but not the hearer is privy. Perhaps more direct evidence in support of this prediction comes from our intuitive understanding of attested usage examples such as (6)-(8). In these cases, it appears clear that the use of the selected numeral is not intended primarily to give rise to a range interpretation but instead to state the existence and value of some critical threshold.

Samaras insists his party can still get more than 150 seats in Parliament to form a government on his own.[1] (6)

[Sachin Tendulkar] has now scored more than 11,953 runs in the five-day game.[2] (7)

The only painting to sell for more than $135 million was, perhaps unsurprisingly, one of Pollock's.[3] (8)

The correct interpretation of these examples seems to rely upon the recognition that the choice of numeral carries additional information, and then a processing of reasoning based on encyclopedic knowledge as to what that information is. In (6), it is that you need 151 seats to command a majority in the Greek parliament; in (7), it is that the record Tendulkar broke previously stood at 11,953 runs; in (8), it is that the second most costly painting ever sold (at that time) cost $135 million. Range interpretations appear to be deprecated here: Samaras is not reported as refusing to assert that he will win 160 seats, and (7) remains both true and appropriate when Tendulkar exceeds 12,000 runs.

In short, while the precise inferences drawn in this case clearly rely on encyclopedic knowledge which is not encoded in these examples, the fact that some type of inference is invited does appear to be a pragmatic property of these examples. If so, this can only be explained under the assumptions that (i) the speaker may choose to

---

1 http://www.ekathimerini.com/4dcgi/_w_articles_wsite1_31219_22/04/2012_438798, retrieved 12 May 2012.

2 http://www.eatsleepsport.com/cricket/tendulkar-breaks-laras-test-record-800854.html#.T67HduVPQms, retrieved 12 May 2012.

3 http://www.dailyiowan.com/2011/02/15/Opinions/21376.html, retrieved 12 May 2012.

use a numeral that is contextually salient and (ii) the hearer is aware of this and interprets the utterance accordingly. To look at it another way, the speaker may exploit the hearer's awareness of how the numerical quantification system works, and use this to signal the significance of some particular numeral, at a certain cost in the informativeness of the utterance with respect to the quantificational task at hand.

It could therefore be argued that this type of speaker-referring approach succeeds in capturing at least some useful generalizations about the interpretation of utterances concerning numerical quantity. However, although the above discussion goes some way towards characterizing the task of the speaker, it does not explain how the speaker is able to perform this task, which is presumed to be a complex matter of abductive inference. In the following section, I turn to the question of heuristics, with a view to outlining some hypotheses about the way in which the hearer actually computes the kind of interpretations that are attested.

5 Constraining the hearer's reasoning

5.1 Motivation

The account discussed above does seem to place very considerable burdens on the hearer's reasoning faculties. As mentioned, Cummins et al. (2012) elicit some evidence via self-report suggesting that hearers do reason in this way. However, such data should be interpreted with caution, in that any attempt to introspect about the workings of such a process might easily result in a *post hoc* rationalization being reported instead. The descriptive utility of the account does not necessarily mean that its internal workings are correct. To draw an analogy with Gricean pragmatics in general, there is widespread acceptance that (something like) Grice's (1975) maxims represent an (approximately) accurate characterization of interlocutor behavior, but debates continue to rage about the mechanisms by which this behavior is brought about. In particular, there is a widely shared intuition that the logical derivation of pragmatic inferences is too laborious to be undertaken in customary practice, and that language users make use of heuristics to short-cut the reasoning process and obtain essentially the same outcomes (on most occasions) at a great saving of time and effort (see for instance Levinson 2000). In that spirit, I consider how the hearer might go about obtaining interpretations of utterances that are approximately correct, according to the model under discussion here, without going through the whole rigmarole of propositional reasoning about the speaker's possible constraint rankings, understanding of prior context, and so on.

A particular concern with the model as articulated above is that the hearer is entitled to assume, on hearing a particular utterance, that any other utterance would have been non-optimal under the prevailing conditions, and to draw pragmatic inferences based on this observation. For this to be a useful capability, in practice, the hearer requires some direction. It would be perfectly valid to infer from the use of "more than 70" that the quantity under discussion was probably less than a million, but intui-

tively this is unlikely to be a useful enrichment of the meaning conveyed[4]. Moreover, this is already entailed by more useful potential enrichments such as "not more than 100". Similarly, it would be valid to infer that an utterance of "more than 40" would have been under-informative, but this adds no information above and beyond the semantic entailment of the utterance. From a Relevance Theory viewpoint, such processing should not take place, on the basis that the effort presumably required in computing the inferences far outweighs any advantage accruing to the hearer from having done so.

Thus, to make this account both tractable and psychologically plausible, it appears to be necessary to posit some conditions that govern the hearer's reasoning. In particular, we wish to constrain the possible alternatives that are considered – or more precisely, to account for how the hearer manages to locate the alternatives that would give rise to pragmatically useful implicatures. In the following subsections I discuss some of the considerations that might be relevant in helping the hearer direct their inferential attention in appropriate directions.

5.2 Exploiting scale granularity

Using the notion of scale granularity, in the sense of Krifka (2009), might expedite the hearer's reasoning process. Many quantity scales have especially salient divisions and units. For the integers, the subset (10, 20, 30…) seems to constitute a natural subscale at a coarser level of granularity than the whole set provides, as does (50, 100, 150…), while for instance (7, 17, 27…) does not. These scales vary across domains. For instance, in time, (15, 30, 45, 60) is a salient scale for minutes, but (3, 6, 9, 12) is one for months, and so on.

Assuming that such scales have some kind of psychological reality, a hearer could exploit these in searching for a particular alternative utterance to use as a basis for pragmatic enrichment. An appropriate heuristic would be as follows: given an utterance, generate an alternative by finding the next scale point that would make this utterance stronger, and (under specific conditions) infer the falsity of this alternative. In the case of expressions such as "more than", or existential contexts, the relevant scale direction is upwards, as in (9) and (10); in the case of "fewer than", it is downwards, as in (11).

Utterance: The child is more than 3 months old.

Next scale point: 6 months.

Alternative: The child is more than 6 months old.

Inference: The child is not more than 6 months old. (9)

[4] We might reasonably suppose that in most circumstances where "more than 70" is asserted, some form of practical upper bound is already common knowledge and need not be communicated. In discussing how many people attended a lecture, based on shared knowledge of lecture venues you might reasonably assume that it was less than 1000, without drawing any inferences from a specific description of the value.

Utterance: We can comfortably accommodate 250 guests.

Next scale point: 300.

Alternative: We can comfortably accommodate 300 guests.

Inference: It is not the case that we can comfortably accommodate 300 guests. (10)

Utterance: There are fewer than 70 places remaining.

Next scale point: 60.

Alternative: There are fewer than 60 places remaining.

Inference: There are not fewer than 60 places remaining. (11)

This type of reasoning is somewhat appealing intuitively, and matches the self-reported behavior mentioned earlier. Within the constraint-based account discussed here, it is also a very efficient approach. The constraints NSAL and GRAN require the use of a number with a particular level of salience in the appropriate context, and disfavor alternatives which use less salient numbers. For instance, in a case such as (9), it would potentially violate GRAN to say "more than 4 months". Consequently, the inference that this claim is false would be unwarranted: the decision to utter (9) might reflect adherence to GRAN rather than any uncertainty on the speaker's part as to the truth of the stronger statement. Thus, with reference to the granularity constraint, the inference based on the next scale point is the first safe inference. It is also the strongest safe inference, in that it entails all the inferences that could have been drawn by considering more distant scale points and the utterances that correspond to them.

5.3 Exploiting prior use of numerals

In the case of numerals which have previously been mentioned in the context, the constraint-based approach predicts that no implicatures of the above kind are robust. In such cases in general, the heuristic proposed in 5.2 should not be applied, on the grounds that it involves appreciable processing effort in the service of a pragmatic enrichment that is conceptually ill-founded. In this circumstance, all stronger alternatives involving different numerals are potentially dispreferred for reasons that have nothing to do with the validity or otherwise of their semantic content.

However, this leaves open the question of how the hearer should act if the numeral might, but might not, be being used because of prior activation. This is arguably the typical case in real life, given that we cannot be certain that we share the speaker's view of the prior context in all relevant particulars, and is also crucial to the interpretation of examples (6)-(8) above. One possibility is that the hearer proceeds to consider alternatives, as specified above, but modulates the confidence associated with the implicature according to how confident the hearer is that the numeral uttered is not contextually salient. That is, a hearer encountering the utterance in (11) should be confident in the corresponding enrichment if they are confident that 70 is not a salient

number in the speaker's view of the prior context. If they feel that 70 might be a salient number, the inference should be drawn with low confidence.

The descriptive accuracy of this account is an empirical question that has not yet been explored. It is worth noting that the participants in Cummins et al.'s (2012) study continued to infer pragmatic upper bounds for "more than 70" etc. even when the numeral was previously mentioned: these bounds were merely weaker and less consistently affirmed than in the case of no prior mention. Perhaps participants know that numeral re-use is a potential reason for the inference to fail, but sometimes risk it anyway. This might indicate that the reasoning process described above continues even when it, logically speaking, should not. Further work might establish whether the inferential processes are default-like or automatic, for instance by exploring the behavior of participants under cognitive load (De Neys and Schaeken 2007, Marty and Chemla 2011).

5.4 Awareness of adjacent scale points

It has also been observed in the literature that expressions such as "more than four" fail to implicate "not more than five", while "at least four" similarly fails to implicate "not at least five" (Krifka 1999, Fox and Hackl 2006). In both cases, the semantic meaning and the implicature would, taken together, entail a precise value ("five" and "four" respectively). This would in turn make the utterance pragmatically odd, given that it was a particularly obscure way to convey this information. However, it does not seem intuitively feasible that the inference should first proceed, and then be cancelled when the hearer realizes that the enriched interpretation is anomalous for the utterance. Furthermore, Cummins et al. (2012) provide evidence that, for instance, "more than 70" does implicate "not more than 80"; and introspectively it seems that "more than 4 meters" does implicate "not more than 5 meters". How is it that hearers refrain from calculating implicatures from "more than four people" but do so for "more than 70 people" or "more than 4 meters"?

One plausible explanation is that the granularity-based heuristic does not operate when scale points are adjacent. This would be a sensible preference from a teleological viewpoint, in that the resulting interpretations would be better expressed by bare numerals. An alternative explanation which does not rely on teleology is that there are in fact contextual factors present in such usage examples which consistently militate against the operation of the heuristic. Specifically, the examples that are discussed in the literature typically appear to involve the repetition of contextually salient values, which according to the account presented here would invoke priming constraints and cause inference failure. For instance, (12) is especially felicitous in a context in which have three children is a critical threshold of some kind, e.g. for the receipt of benefits, that has already been established in the discourse. This intuition appears to be widespread, but its potential pragmatic significance appears not to have been discussed. Once again, empirical work could establish whether these explanations could be tenable.

John has more than three children. (12)

5.5 Looking for simpler quantifiers

An interesting question concerns the relations that hold between different quantifiers. When we consider alternative expressions that involve the same number, such as "at least four" and "more than four", clearly the number-referring constraints cannot adjudicate between them: both expressions incur the same number of violations of NSAL, NPRI and indeed GRAN. The relevant constraints are therefore quantifier simplicity (QSIMP), quantifier priming (QPRI) and potentially informativeness (INFO). Considering first the situation in which no (quantifier) priming is in effect, the hearer should be entitled to conclude that the speaker could not use a simpler quantifier and the same number.

To make any theoretical use of this observation, we have to be able to establish what constitutes a 'simpler' quantifier. A bare numeral (i.e. one with a null quantifier) could plausibly be argued to use the simplest possible quantifier, but precedence among other expressions is not so accessible to introspection. Nevertheless, Cummins and Katsos (2010) argue from experimental data that the superlative quantifiers 'at least' and 'at most' are more complex than the corresponding comparative quantifiers 'more than' and 'fewer than'. Thus, we can say of 'at least four' that it is more complex than 'four' and more complex than 'more than four'; similarly, *mutatis mutandis*, for 'at most four'. If this conclusion is correct, the model would predict that the inferences specified in (13) and (14) are legitimate for the hearer in the absence of a contextually primed quantifier.

John has at least four children.

Alternative 1: John has four children.

Inference 1: The speaker is not certain that John has (exactly) four children.

Alternative 2: John has more than four children.

Inference 2: The speaker is not certain that John has more than four children. (13)

John has at most four children.

Alternative 1: John has four children.

Inference 1: The speaker is not certain that John has (exactly) four children.

Alternative 2: John has fewer than four children.

Inference 2: The speaker is not certain that John has fewer than four children. (14)

In both cases, these are weak implicatures (concerned with the speaker's inability or unwillingness to make stronger statements, rather than necessarily the truth or falsity of these stronger statements). Taken together, they entail that the speaker of "more than four" (or "fewer than four") considers it possible, but not certain, that 'exactly four' is the case.

By contrast, (15) and (16) admit no such inferences. Holding the numeral constant, the alternatives involving "at least/most" would be more complex and less informative, and are ruled out as a basis for implicature for both reasons. The bare numeral would give an expression ("(exactly) four") that is already contradicted by the semantics of the original statement.

John has more than four children. (15)

John has fewer than four children. (16)

The generalization is that the inferences from "more than *n-1*" are systematically different from those arising from "at least *n*", which coheres with the observation of Geurts and Nouwen (2007). However, unlike their approach, the model presented here generates this prediction without positing any semantic differences between the two expressions. Instead, the difference arises because of the dissimilarity of the set of alternatives for the two expressions. The keystone of this account is in fact the presence of numeral-referring constraints in the system. Given such constraints, the hearer cannot readily draw inferences about the falsity of alternatives involving different numerals to the one actually uttered, because the specific choice of numeral might reflect the operation of a priming constraint. Thus, even if we consider "more than three" and "at least four" to be logically equivalent, the speaker who uses "at least four" signals an unwillingness to assert "more than four", while the speaker who uses "more than three" does not signal an unwillingness to assert "more than four". This offers a possible account of the otherwise problematic asymmetry between two forms which can be argued to convey the same semantic meaning ("greater than or equal to 3") but give rise to different inference patterns (Geurts et al. 2010).

A tacit assumption in the foregoing is that the use of a particular numeral is more likely to be pragmatically determined than the use of a particular quantifier. Were this not the case, then the use of "more than four" would indeed reliably signal an unwillingness to commit to "more than five", which satisfies the same quantifier-referring constraints. There are good reasons to suppose that this is a reasonable assumption, outside of the observation that the above implicature is not considered to be conveyed. First, there is an additional numeral-referring constraint in the posited system, namely GRAN, which also makes stipulations about the numeral that must occur in an alternative expression for it to be the basis for this form of pragmatic enrichment. Secondly, numerals appear to be used more often than their (non-null) quantifier complements, which would mean that numerals are exerting priming effects in more actual contexts than quantifiers are.

Under these assumptions, a sensible heuristic would be to consider less complex alternative quantifiers. A first simple alternative might be the null quantifier: that is, the hearer may respond to a quantifier before a numeral by inferring that the speaker cannot commit to the corresponding expression with the quantifier omitted. In the case of "more than", this will yield a pragmatic inference that is already entailed by the semantic meaning, but in other cases ("at least", "at most", "about", "no less than", and so on) such an inference would augment the semantic meaning, if we assume a traditional mathematical/set-theoretical analysis for the semantics in question.

The heuristic might then proceed to consider "more/fewer than" as an alternative. Cummins and Katsos (2010) argue empirically that these expressions are less complex than "at least/most", also drawing on observations marshaled by Geurts and Nouwen (2007). In particular, the former paper hypothesizes that the relation of strict comparison is simpler than the relation of non-strict comparison (i.e. greater/less than or equal to). If correct, this claim further predicts that all the class B quantifiers of Nouwen (2010) are more complex than "more/fewer than". The implications of this for "at least/most" are sketched out above, and a similar story can be told for "no more/fewer than". However, if "more/fewer than" represents a particularly simple mode of comparison, it could conceivably even enter into consideration as an alternative for expressions such "about/around/approximately *n*". This would give rise to the prediction, for instance, that the use of such an approximation (at least weakly) implicated that the speaker did not know whether the value was above or below (or equal to) *n*. This appears to be a plausible claim but one that presently lacks empirical support, as does the more general hypothesis about the absolute simplicity of "more/fewer than" that underpins it.

A separate possibility is that the alternatives to be considered are selected on the basis of semantic or associative proximity to the forms that were actually uttered. If we adopt a spreading activation model, the presumed fact that the hearer of "at least" considers "more than" as an alternative could be attributed to three distinct causes. One is the simplicity of "more than", as argued above; a second is the closeness of the relation between the semantic meanings of "at least" and "more than", in a formal sense; and a third is the closeness of this relation in a spatial sense (for instance, in the similarity of the expressions' distribution patterns, à la Landauer and Dumais 1997). There is experimental evidence that the accessibility of alternatives is highly correlated with the availability of implicatures about their falsity (Zevakhina 2011). However, whether this accessibility is itself due to distributional similarity, conceptual proximity, or merely general familiarity, is a matter for future experimental and theoretical research.

5.6 Potential implications for pragmatics

How does this account compare to standard pragmatic approaches? Its major stipulation is that certain specific alternatives are considered in place of the material that was actually uttered (and consequently, that the process determining which alternatives are considered is pertinent to determining which inferences are derived). This could be seen as occupying an intermediate position between default and contextual accounts of implicature. In common with default accounts, this approach suggests that certain expressions serve as triggers for inferences about the falsity of other specific expressions. However, in common with contextual accounts, it predicts that implicatures are not derived in cases where their licensing conditions are not met. The justification for positing the partial element of defaultness in this model is practical, in that it is not feasible for the hearer to evaluate all the possible alternatives, and the search-space for pragmatically significant alternatives must therefore be curtailed in some way. However, I propose that these potential alternatives are then evaluated on their

merits, and that there is no rush to judgment about their falsity. This approach respects the intuitions of theorists such as Levinson (2000), while remaining compatible with the major findings of the experimental pragmatic literature, which has so far uncovered very little evidence for the presence of true default inferences.

6 Conclusion

A unidirectional OT approach to numerical quantification yields novel predictions as to the usage and interpretation of such expressions. Unlike other approaches, it assumes that speaker meaning is not fully and accurately recovered by the hearer in general. Aspects both of hearer behaviour and of their introspective experience appear to support this general view. Nevertheless, it places a considerable burden of reasoning on the hearer, requiring a practical account of how the pragmatic task can be rendered tractable. This paper sketches part of such an account and explores its implications, which notably include further predictions about the interpretation of expressions and about their relative complexity, which could usefully be the object of further empirical investigation.

References

1. Blutner, R. (2000): Some Aspects of Optimality in Natural Language Interpretation. Journal of Semantics 17, 189–216.
2. Blutner, R. (2006): Embedded Implicatures and Optimality Theoretic Pragmatics. In: Solstad, T., Grønn, A. and Haug, D. (eds.) A Festschrift for Kjell Johan Sæbø: in partial fulfilment of the requirements for the celebration of his 50th birthday. Oslo.
3. Bonnefon, J. F., Feeney, A. and Villejoubert, G. (2009): When Some is Actually All: Scalar Inferences in Face-Threatening Contexts. Cognition 112, 249–258.
4. Breheny, R. E. T., Ferguson, H. J. and Katsos, N. (2013): Taking the Epistemic Step: Toward a Model of On-line Access to Conversational Implicatures. Cognition 126, 423-440.
5. Büring, D. (2008): The Least 'At Least' Can Do. In: Chang, C. B. and Haynie, H. J. (eds.) Proceedings of the 26th West Coast Conference on Formal Linguistics, pp. 114–120. Cascadilla Press, Somerville MA.
6. Cummins, C. (2011): The Interpretation and Use of Numerically-Quantified Expressions. PhD thesis, available online at http://www.dspace.cam.ac.uk/handle/1810/241034.
7. Cummins, C. and Katsos, N. (2010): Comparative and Superlative Quantifiers: Pragmatic Effects of Comparison Type. Journal of Semantics 27, 271–305.
8. Cummins, C., Sauerland, U. and Solt, S. (2012): Granularity and Scalar Implicature in Numerical Expressions. Linguistics and Philosophy 35, 135–169.
9. De Neys, W. and Schaeken, W. (2007): When People are More Logical Under Cognitive Load: Dual Task Impact on Scalar Implicature. Experimental Psychology 54, 128–133.
10. Fox, D. and Hackl, M. (2006): The Universal Density of Measurement. Linguistics and Philosophy 29, 537–586.
11. Geurts, B., Katsos, N., Cummins, C., Moons, J. and Noordman, L. (2010): Scalar Quantifiers: Logic, Acquisition, and Processing. Language and Cognitive Processes 25, 130–148.
12. Geurts, B. and Nouwen, R. (2007): "At least" et al.: the Semantics of Scalar Modifiers. Language 83, 533–559.

13. Grice, H. P. (1975): Logic and Conversation. In: Cole, P. and Morgan, J. L. (eds.) Syntax and Semantics, Vol. 3. Academic Press, New York.
14. Hendriks, P. and de Hoop, H. (2001): Optimality Theoretic Semantics. Linguistics and Philosophy 24, 1–32.
15. Horn, L. R. (1984): Towards a New Taxonomy for Pragmatic Inference: Q-based and R-based Implicature. In:Schiffrin, D. (ed.) Meaning, Form and Use in Context (GURT '84), pp.11–42. Georgetown University Press, Washington DC.
16. Krifka, M. (1999): At least some Determiners aren't Determiners. In: Turner, K. (ed.), The Semantics/Pragmatics Interface from Different Points of View. Current Research in the Semantics/Pragmatics Interface Vol. 1, pp.257–292.
17. Krifka, M. (2002): Be Brief and Vague! And how Bidirectional Optimality Theory allows for Verbosity and Precision. In: Restle, D. and Zaefferer, D. (eds.) Sounds and Systems. Studies in Structure and Change. A Festschrift for Theo Vennemann, pp.439–458. Mouton de Gruyter, Berlin.
18. Krifka, M. (2009): Approximate Interpretations of Number Words: a Case for Strategic Communication. In: Hinrichs, E. and Nerbonne, J. (eds.) Theory and Evidence in Semantics, pp.109–132. CSLI Publications, Stanford.
19. Landauer, T. K. and Dumais, S. D. (1997): A Solution to Plato's Problem: the Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. Psychological Review 104, 211–240.
20. Levinson, S. C. (2000): Presumptive Meanings. MIT Press, Cambridge MA.
21. Marty, P. and Chemla, E. (2011): Scalar Implicatures: Working Memory and a Comparison with 'Only'. Ms., LSCP.
22. Nouwen, R. (2010): Two Kinds of Modified Numerals. Semantics and Pragmatics 3, 1–41.
23. Smolensky, P. (1996): On the Comprehension/Production Dilemma in Child Language. Linguistic Inquiry 27, 720–731.
24. Sperber and Wilson (1986/1995): Relevance: Communication and Cognition. Blackwell, Oxford.
25. Zevakhina, N. (2011): Strength and Similarity of Scalar Alternatives. In: Aguilar Guevara, A., Chernilovskaya, A. and Nouwen, R. (eds.) Proceedings of Sinn und Bedeutung 16, Vol. 2, pp. 647–658. MIT Working Papers in Linguistics, Cambridge MA.