Modelling context within a constraint-based account of quantifier usage

Chris Cummins[1, 2] and Napoleon Katsos[1]

[1] Department of Theoretical and Applied Linguistics, University of Cambridge

[2] SFB 673 – Alignment in Communication, Bielefeld University

Abstract

Recent work on numerically-quantified expressions has aimed to identify which components of their meaning are semantic and which are pragmatic. Pragmatically-oriented accounts assign a crucial role to contextual factors, such as the level of information requested in the preceding discourse and the availability of certain expressions to the speaker at the time of utterance. However, these models are typically imprecise as to which factors are relevant and how they interact. We discuss a recent proposal that treats numerical quantifier usage as a problem of multiple constraint satisfaction, and demonstrate its descriptive utility. In particular, we consider how the context-referring constraints within this model (governing prior mention of the numeral and the quantifier, granularity level, and informativeness) constitute a proposed definition of relevant context. In such a model, contextual factors can influence the output only if they are referred to in constraint definitions. Therefore, if a model of this type succeeds in capturing the meaning of use and numerical expressions, its constraint definitions specify which contextual factors are relevant to the choice of expression. We consider to what extent this account succeeds in modelling actual speaker behaviour, and whether it generalises to other domains of usage.

1. Introduction

Much recent work on numerically-quantified expressions has aimed to identify which components of their meaning are semantic and which are pragmatic. Drawing upon the work of Barwise and Cooper (1981), researchers have typically assumed that the meaning of such expressions is appropriately analysed in terms of mathematical operators. However, for various classes of expression, these accounts have been shown to be descriptively inadequate.

An example of this is the analysis of superlative quantifiers – those of the form 'at least *n*', 'at most *n*'. On a traditional account, these might be assumed to be semantically equivalent to the mathematical formalisms '$\geq n$' and '$\leq n$' respectively. However, Geurts and Nouwen (2007) demonstrate that superlative quantifiers cannot simply be treated as encodings of these operators. With respect to cardinal quantities, this formalism would imply that 'at least *n*' was equivalent to 'more than *n*-1' and 'at most *n*' to 'fewer than *n*+1': that is, superlative quantifiers should be expressible in terms of comparative quantifiers. Geurts and Nouwen (2007) show that superlative quantifiers differ from the corresponding comparative quantifiers in terms of the inferences that they license, and in terms of their distribution.

Motivated by these observations, Geurts and Nouwen (2007) propose an alternative semantic account of superlative quantifiers that adds modality to the quantitative meaning. On their account, 'at least *n*' is argued to convey the certainty of the existence of a group of size *n* and the possibility that more than *n* might be the case; a similar account is offered for 'at most *n*'. Geurts et al. (2010) obtain experimental evidence to support this account above the traditional model. By contrast, Cummins and Katsos (2010) offer a pragmatic account of superlative quantifier meaning, motivated by the experimental finding that non-strict comparison ('greater/less than or equal to') is more costly than strict comparison ('greater/less than') in terms of verification time, even when the comparison is encoded symbolically rather than in

words. They argue that the modal meaning identified by Geurts and Nouwen (2007) arises through implicature as a result of this difference in processability, and present further empirical evidence in support of this contention. In particular, the modal meaning is shown to arise only in contexts where it would make the utterance more informative, such as declaratives, and does not arise in other contexts, such as the antecedents of conditionals.

Pragmatically-oriented accounts of numerical quantification assign a crucial role to contextual factors in determining the usage and meaning of utterances. This raises the question of precisely which factors are responsible and how they interact. For instance, Van der Henst, Carles and Sperber (2002) discuss time-reporting from a relevance theory perspective. In their experiments they asked participants for the time in two distinct conditions, a neutral condition and an exact condition in which the requester implicitly expresses an interest in the exact time (saying that they wish to set their watch). They demonstrate that participants are more likely to give exact times if they wear digital rather than analogue watches, and in the exact rather than the neutral condition.

Van der Henst and Sperber (2004) argue that this study provides strong evidence that the speaker's choice of expression involves the interaction of distinct factors. Rounding is argued to enable both speaker and hearer to work with values that are more cognitively salient (those corresponding to marks on the clock-face). It also reduces the speaker's commitment and might be favoured if the speaker doubts the precision of their information. They consider these factors to be contributory to relevance. In addition to these general dispositions, however, they also argue for the role of contextual factors in determining the speaker's choice of utterance. First, they assume that wearers of digital watches should tend to give precise answers, presumably because these answers are literally spelled out in front of the speaker. Moreover, they demonstrate that speakers respond to their interlocutors' cues as to the precision of the information required, by showing that rounded answers occur less

frequently in the exact condition than in the neutral condition, as described above. This information constitutes part of the discourse context as far as the speaker is concerned.

Thus, Van der Henst and Sperber (2004) argue that several distinct factors interact in determining the preferred utterance in a given situation, including factors that we can clearly regard as part of the speaker's context. However, their account fails to specify how these contributory factors interact. This omission is significant given that it is typically impossible for the speaker to produce an utterance that is wholly satisfactory with respect to all criteria. For instance, in the time-reporting case, the speaker is forced to choose between giving a maximally precise answer and giving one that is rounded to a convenient level of precision. Both of these would presumably be desirable attributes from a relevance perspective, but the utterance typically cannot possess both. Accepting the general assumptions of Van der Henst and Sperber, among others, the speaker's choice of utterance must therefore be presumed to reflect a decision as to which attributes are most important.

Cummins (2011) attempts to address this omission by proposing that the speaker's behaviour can be modelled as a process of multiple constraint satisfaction (within an Optimality Theory framework). In this model, optimal outputs are those which best satisfy a ranked set of constraints. These constraints govern both the complexity of the utterance and its relation to the prevailing 'situation' at the time of utterance. In this way, the constraint-based approach offers a way of integrating contextual information into the speaker's decision procedure. In particular, the model posits a role for informativeness, which relates the utterance to the knowledge states of the interlocutors; granularity, which relates the utterance to the level of precision called for by the discourse; and prior activation of the numeral and the quantifier, which relate the utterance to the co-text (or potentially to the external environment).

In this chapter, we discuss the merits of such an account as a method for modelling context. We consider first how this model draws a distinction between overtly contextual considerations and those which are less situation-dependent. We briefly review the specific constraints proposed by Cummins (2011) and consider alternative ways of treating the behaviour that these constraints are intended to model. Then we consider how this gives rise to a proposed definition of relevant context, and evaluate its plausibility. Finally, we turn to the hearer's role within such a system, examining the interplay between contextual factors and pragmatic enrichments and discussing experimental data that suggests this is a two-way process for the hearer.

2. Context-referring and context-independent constraints

The approach adopted by Cummins (2011) treats the factors influencing the speaker's choice of utterance as 'soft' or violable constraints. Specifically, these constraints are then analysed within an Optimality Theory (OT) framework. The precise details of this account are not germane here, as we shall discuss later: we do, however, wish to exploit a distinction identified within OT, namely that between markedness and faithfulness constraints (Prince and Smolensky 1993).

Within OT, markedness constraints govern the form of the output, irrespective of the input, while faithfulness constraints govern the relation between the input and output. A phonological example of a markedness constraint is NOCODA or -COD (Prince and Smolensky 1993: 93), which is violated by syllables with codas. As for any markedness constraint, whether an output candidate violates this constraint can be determined just by inspecting that output candidate. A phonological example of a faithfulness constraint is DEP, 'don't epenthesise' (McCarthy 2002: 13), which specifies that segments should not be present

in the output that are absent from the input. As for any faithfulness constraint, it is necessary to compare the input with the output candidate in order to determine whether DEP has been violated.

Within the domain of phonology, OT has been employed to model the mapping between underlying forms and surface forms, the former being treated as inputs for the speaker and the latter as outputs. In pragmatics, early applications of OT (Hendriks and de Hoop 2001 i.a.) took the hearer's perspective and aimed to model the interpretation of utterances. From this perspective the utterance can be considered as the input and the interpretation as the output. The proposal of Cummins (2011) takes the speaker's perspective, and thus reverses this pattern, treating the situation (including the speaker's intention) as the input and the utterance as the output. Work using bidirectional OT (Blutner 2006 i.a., Krifka 2009) aims to model optimal form-interpretation pairs, and hence can be regarded as optimising from both the speaker's and the hearer's perspective simultaneously. Within this model, the utterance is both an input and an output, as is the interpretation.

However, the distinction between markedness and faithfulness constraints is manifest whether we consider the utterance to be the input and the interpretation the output or vice versa. Crucially, faithfulness constraints relate the two levels (form and meaning), while markedness constraints are restricted to one level. For example, Krifka (2009) proposes two markedness constraints, SIMPEXP (which favours simple rather than complex expressions) and APPRINT (which favours approximate rather than precise interpretations). As markedness constraints, these function at only one level: SIMPEXP favours a simple expression irrespective of the meaning to be conveyed, and APPRINT favours an approximate interpretation irrespective of the expression used. In his bidirectional model, Krifka uses these to recover what he calls the RNRI principle (Krifka 2002), which observes

that round numbers tend to have round interpretations. If this principle were to be encoded as a single constraint, it would necessarily be a faithfulness constraint, because it specifies an interpretation conditional upon a property of the utterance. That is, a putative RNRI constraint would favour round interpretations when the utterance used a round number and non-round interpretations when the utterance did not[1]. Violations of this RNRI constraint could only be determined with reference to both the proposed interpretation and the utterance that took place.

If we are to model context by the use of OT-type constraints, it therefore follows that these must be faithfulness constraints. Markedness constraints are context-blind. In the hearer-referring model of Hendriks and de Hoop (2001), markedness constraints effectively state that certain kinds of interpretation are preferred irrespective of the utterance that is being interpreted. In the speaker-referring model of Cummins (2011), markedness constraints effectively state that certain kinds of utterance are preferred irrespective of the situation that is being described. Both are intuitively valid observations, suggesting that there is a place for markedness constraints within such a model. However, it is clear that unidirectional OT models must handle contextual factors by appeal to faithfulness constraints if they are to do so at all[2].

[1] Krifka (2009: 109) does not consider modelling the RNRI principle in this way, presumably because he considers it "hard to imagine that it is an irreducible axiom of language use". Nevertheless, it seems plausible that the RNRI is directly functionally motivated by virtue of round numbers corresponding to scale points on an accumulator scale (Dehaene 1997). Such an approach would provide a possible account of the processing of round numbers that is not attempted by Krifka (2009), who aims instead to characterise the preferred form-meaning pairs in the system as a whole.

[2] The argument as applied to bidirectional models is not so straightforward. Nevertheless, we hold the view that modelling contextual factors in a bidirectional approach presents serious difficulties: for instance, the question arises of whether the same meaning in two different contexts necessarily requires two different forms of expression, in order that the form-meaning pairs should not overlap. Given these issues and questions about the psychological plausibility of bidirectional OT as a processing model (Blutner 2006), we do not consider it in detail here.

3. Constraints proposed to model context

In the previous section we argued that faithfulness constraints would be required to treat context. In this section we exemplify this, focusing initially on the speaker-referring proposal of Cummins (2011). We discuss its faithfulness constraints and consider the extent to which these can be considered to encode aspects of context.

Cummins (2011) posits six constraints: informativeness, granularity, quantifier simplicity, numeral salience, quantifier priming and numeral priming. Of these, quantifier simplicity and numeral salience are markedness constraints and are independent of context, as discussed above. The four remaining constraints are faithfulness constraints governing the relation between situation and utterance. However, the notion of 'situation' within this model is very broad, encompassing in principle any information that is available to the speaker, including aspects of the speaker's psychological state as well as such things as the contents of the preceding dialogue and the visual scene. Correspondingly, we can distinguish between the aspects of context modelled by the individual faithfulness constraints, which vary from linguistic to perceptual to psychological considerations.

First, the constraint on informativeness is defined in terms of the speaker's intention and requires the speaker to make the most informative statement available about the topic under discussion. With respect to this constraint, the violations incurred by a candidate output depend upon the epistemic state of the speaker: "more than 90" would incur a violation if the speaker knows that "more than 91" holds, but would not incur a violation extended only to "more than 90". We could consider the speaker's knowledge to be an aspect of the context, but considerations of epistemic state have generally been treated separately in pragmatic research, so we leave this possibility aside in this chapter.

The constraint on granularity encodes more traditional contextual considerations. Granularity refers to the density of representation points on scales: for instance, the major representation points associated with time include those denoting minutes, 5-minute periods, quarter-hours and hours. The constraint is violated by a failure to use the level of granularity required by the context, and thus assumes a role for the context in specifying this parameter. As an example of how this would work, we can revisit the research of Van der Henst, Carles and Sperber (2002) on the reporting of time. They demonstrate that participants who are asked for the time respond with a precise answer more frequently when the asker implicitly specifies a request for precise information – i.e. when the asker expresses a wish to set their watch. In their terms, this is taken to indicate that the respondent is concerned with providing the most relevant information, but they do not explain how the speaker determines how to do this. We might therefore argue that the speaker's behaviour arises in accordance with a desire to provide information at the appropriate granularity level, where this is specified by the preceding discourse context. Moreover, the fact that their participants did not always adhere to this requirement, apparently contrary to the expectations of relevance, is evidence in favour of the treatment of granularity as a violable constraint.

The constraints on numeral priming and quantifier priming also encode contextual factors. The numeral priming constraint is violated by a failure to reuse a numeral that has become activated in the preceding discourse, or which is salient in the speaker's perceptual environment (if such a numeral exists). It is motivated by appeal to the notion that an activated numeral of this kind will be more accessible for subsequent use by the speaker, and thus yield priming effects (in the broad sense of Pickering and Garrod 2004). Similarly, the constraint on quantifier priming is violated by a failure to reuse a quantifier that has become activated in the preceding discourse, in cases where this has occurred. The prior use of a quantifier is similarly argued to yield priming effects, as well as potentially giving rise to

syntactic alignment. Both constraints permit contextual factors related to the discourse and environment to exert influence upon the speaker's choice of expression.

Within the model, there is an implicit division of labour between the above priming constraints and the corresponding markedness constraints. This is especially striking in the case of numeral priming and numeral salience. The latter constraint assumes that round numbers are generally preferred on psychological grounds (Dehaene 1997) and is violated by the failure to use a round number. Meanwhile, the former constraint assumes that contextual considerations influence which number is preferred in a given situation. Thus, Cummins (2011) supposes that the choice of numeral is evaluated with respect to two distinct considerations within this model – its contextual salience and its absolute salience. Similarly, the choice of quantifier is argued to be evaluated with respect to its contextual activation, on the one hand, and its general simplicity, on the other.

This approach is theoretically tenable and maintains clarity as to the precise role of context, but is arguably counter-intuitive. We might alternatively propose that the availability of a number to a speaker, at any given moment, is determined both by the immediately preceding context and the sum total of the speaker's previous experience of that number. Such an analysis would conflate contextual and absolute considerations of salience and propose that both are evaluated simultaneously. Indeed, from this point of view, it could be argued that the general 'landscape' of numeral salience in the mind of the speaker is itself an aspect of context. A similar argument can be made for the case of quantifiers. Ultimately the precise way in which the constraints interact in determining the utterance is not crucial to the matter at issue here, as we discuss in section 5.2, but does raise questions about the possibility or desirability of distinguishing external context from aspects of the speaker's psychological state.

In the following subsections we consider how these constraints explain some observations about the effect of context on the usage and interpretation of numerical quantifiers. In the subsequent section we will then consider alternative context-referring constraints, and explore whether these merit addition to the system, before exploring how the constraint set can be restated as a proposal about the nature of relevant context.

3.1. Comparative quantifier implicatures and priming effects

Cummins, Sauerland and Solt (submitted) investigate the effect of granularity and numeral priming on the range of interpretation of numerical expressions. Specifically, they show that the use of comparative quantifiers such as "more than *n*" gives rise to scalar implicatures, and that these are conditioned by granularity considerations and by prior mention of the numeral. In a series of experiments, they ask participants to read dialogues in which these quantified expressions occur, and in which the speaker is stated to be informed and cooperative. They then ask participants to specify either the range of values that they feel the speaker is referring to, or the most likely value. An example dialogue is given in (1).

(1) A. This display case holds CDs. How many CDs do you own?

B. I have more than 60 CDs.

Their results show that the use of "more than *n*" for round *n* gives rise to implicatures that "more than *m*" does not hold, for certain values of *m*. Specifically, *m* must be greater than *n* and must be a scale point of some scale of coarser, or equally coarse, granularity than the scale on which *n* is a scale point. For instance, "more than 70" tends to implicate "not more than 80", "more than 80" tends to implicate "not more than 100", and "more than 93" tends to implicate "not more than 100". This demonstrates awareness either of granularity or

numeral salience considerations in the interpretation of utterances: hearers assume that the speaker wishes to use a salient numeral, or an expression at a particular granularity level, and thus only draw implicatures about the falsity of other candidate utterances involving numerals matched in salience or granularity to those that were in fact uttered.

The above finding is ambiguous: it could reflect awareness of contextual considerations (the need to use a number of appropriate granularity) or general psychological factors not mediated by context (the need to use a number that is high in salience). However, Cummins et al. go on to demonstrate an unambiguously contextual effect, namely that the prior mention of a numeral gives rise to weaker implicatures. They construct minimal pairs of dialogues in which a numeral is or is not previously mentioned: for instance, (2) is the dual to (1) above. They then compare participants' interpretations of utterances in dialogues of these two types.

(2) A. This display case holds 60 CDs. How many CDs do you have?

B. I have more than 60 CDs.

They demonstrate that participants draw weaker implicatures in the 'primed' than in the 'unprimed' case: that is, they consider B's utterance in (2) to be compatible with a significantly greater range of interpretation than B's utterance in (1). The difference can be accounted for in terms of numeral priming. According to this account, B's utterance in (1) is assumed to be the most informative that uses a highly round number. "More than 80" would be equally satisfactory in this respect, so B's failure to use it tends to suggest that it would not be true. By contrast, B's utterance in (2) might reflect a desire to reuse the number used by A, which would in this case presumably make it easier to draw the intended inference about B's attitude to the display case (i.e. that it is not big enough). "More than 80" would not be as satisfactory in this regard, because it violates numeral priming: therefore the failure of B to use this statement does not strongly implicate that it does not hold. Thus, just as the

granularity and numeral salience constraints of Cummins (2011) predict the availability of the implicated upper bound for "more than *n*" (and similarly, an implicated lower bound for "fewer than *n*"), so the numeral priming constraint correctly predicts the attenuation of these bounds. However, it should be noted that this effect might alternatively be attributed to informativeness considerations, as we consider in section 4.1.

3.2. Use of comparative quantifiers with non-round numerals

Given the above findings, it is not immediately apparent why "more than *n*" is ever admissible for non-round *n*. According to this account, a knowledgeable speaker who utters (3) should implicate (4), and these together entail (5).

(3) The casino uses a pack of more than 52 cards for blackjack.

(4) The casino does not use a pack of more than 53 cards for blackjack.

(5) The casino uses a pack of 53 cards for blackjack.

This is clearly not the intended meaning of (3), and in any case a speaker who wished to convey this meaning would presumably be better off saying (5) than saying the more verbose (3) and relying on the hearer to draw the above inferences.

So, why does (3) not implicate (4)? One explanation is that "more than *n*" systematically fails to yield implicatures; but the results of Cummins et al. (submitted) strongly suggest that this is not generally true.

By appealing to numeral priming, we can propose a more satisfactory explanation. According to this account, usages such as (3) arise because the speaker is adhering to a priming constraint requiring that the numeral (in this case, 52) should be used. This could

arise if the numeral is contextually highly activated, a condition which is met in this case because 52 is an especially salient number in the context of card games (namely, the size of a standard pack of cards). In this case, the implicature process is blocked, just as in the previous subsection: the hearer is aware that the speaker may have chosen to make an informationally weaker statement in order to satisfy numeral priming.

Examples of this are apparently widespread, while examples of "more than *n*" for non-round and contextually non-salient *n* seem to be rare. An online search turned up the following examples: (6) depends upon the fact that the current record for golf major wins is 18, (7) refers to the number of days in a year, and (8) refers to the speed of sound in miles per hour.

(6) Will Tiger win more than 18 majors?[3]

(7) Only studies reporting outcomes after more than 365 days were selected[4].

(8) It will take a lot more than 768 mph to leave the Earth's gravitational hold[5].

In all these cases, the use of a specific number as a reference point appears to be motivated by the salience of this number in the preceding context. As a consequence, the natural interpretation of these quantifiers appears to correspond closely to their semantics, on a traditional view. Thus, in this case, the numeral salience constraint explains how contextual factors prevent pragmatic enrichments from proceeding.

3.3. Non-isomorphic readings of quantifiers

The quantifier priming constraint can be used to account for speakers' behaviour when correcting quantified statements. This encompasses behaviour that might conventionally be

[3] http://www.sporttaco.com/rec.sport.golf/Will_Tiger_win_more_than_18_majors_3677.html, retrieved 24 July 2011

[4] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2780848/, retrieved 24 July 2011

[5] http://www.winnipesaukee.com/forums/archive/index.php/t-10499.html, retrieved 24 July 2011.

treated as metalinguistic. In particular, it could be used to account for the availability of non-isomorphic usages and interpretations, in the sense of Musolino, Crain and Thornton (2000).

Musolino's Observation of Isomorphism states that semantic scope assignments preferentially correspond to syntactic scope. For instance, (9) is ambiguous between readings (10) and (11), but (10) is preferred (in which the quantifier outscopes the negation).

(9) All of the toys are not in the box.

(10) None of the toys are in the box.

(11) Some (but not all) of the toys are in the box.

Intuitively it appears that the non-isomorphic reading is most naturally obtainable when (9) occurs in the context of a prior mention of "all of the toys". The quantifier priming constraint can account for this observation. Suppose a speaker wishes to convey the meaning of (11). In the absence of a preceding context, this could efficiently be accomplished by uttering (11). However, given a preceding context such as (12), the quantifier priming constraint might militate in favour of the utterance of (9) rather than (11) as a way of conveying this meaning.

(12) Are all of the toys in the box?

As a consequence of this constraint, the utterance of (9) in response to (12) would be expected to evoke the meaning (11). By contrast, the utterance of (9) in an unprimed context would not evoke this meaning – indeed, the hearer might legitimately expect that the speaker does not mean (11), as otherwise they would have said (11). Similarly, the utterance (13) might be taken to convey the meaning (11) generally, but to mean (10) when made as a response to (14).

(13) Some of the toys are not in the box.

(14) Are some of the toys in the box?

In this respect, the quantifier priming constraint yields predictions that are coherent with the observation of isomorphism and with our intuitions about when this generalisation fails.

4. Alternative context-referring constraints

In the preceding subsections, we reviewed some of the ways in which the proposed set of context-referring constraints accounts for experimental findings and intuitions about the usage and interpretation of some categories of quantified expressions. The constraints proposed by Cummins (2011) are individually experimentally motivated, in accordance with the discovery procedure for OT constraints laid out by McCarthy (2002: 41f). However, this does not necessarily imply that these represent an optimal set of constraints. In this section we discuss alternative constraints that might be used in the modelling of context effects.

4.1. Question under discussion

Given that the speaker aims to produce an utterance that is helpful for the purpose of the communicative exchange – in other words, that the speaker intends to be relevant in the sense of Grice (1975) – it seems reasonable to suggest that the utterance should be addressed to the question under discussion (Roberts 1996). We could posit a faithfulness constraint requiring that this is upheld.

Such an approach would provide a different treatment of priming effects. For instance, if we revisit example (2) from section 3.1, repeated as (15) below, we could analyse B's decision to reuse the numeral as an attempt to address the implicit question in A's utterance (whether the

display case is suitable for B's needs). It would again follow that the comparative quantifier used by B did not license an implicature, because it had been chosen merely to address A's question in the most direct way possible.

(15) A. This display case holds 60 CDs. How many CDs do you have?

B. I have more than 60 CDs.

Similarly, we could treat the examples in section 3.3 as cases where the utterance is interpreted differently according to whether or not it directly addresses a question under discussion. This resembles Gualmini's (2004) account of the derivation of non-isomorphic meanings. On this view, given (16) as a preceding context, (17) can obtain a non-isomorphic interpretation, because (16) specifies that the question under discussion is whether all the toys are in the box, and even the non-isomorphic interpretation of (17) suffices to answer this in the negative. By contrast, in the absence of a preceding question such as (16), (17) appears to stand as an answer to a less specific question about the location of the toys in general, and thus is preferentially interpreted as predicating a location for all the toys (i.e. isomorphically).

(16) Are all of the toys in the box?

(17) All of the toys are not in the box.

Therefore it appears that the use of a 'question under discussion' constraint would crosscut the application of existing constraints and indeed raise questions about their necessity to the system. However, problems arise in the formulation of such a constraint, most strikingly with how we define the notion of 'question under discussion'. We cannot follow the approach of Zondervan (2007) and identify the question under discussion as the question that the utterance is supposed to answer, because then any utterance would automatically satisfy the constraint and it would thus have no explanatory power. Rather, we need to be able to

identify the question under discussion at any point of the discourse, prior to the selection of the utterance that may (or may not) answer it. This may not be an insuperable problem – speakers seem generally to be adept at guessing what information, out of the vast resources the speaker possesses, will be of particular interest to the hearer at a given moment – but it is not immediately clear how such a constraint can be economically expressed, nor is it clear what constitutes a violation of such a constraint. Future work in this area may shed light on how such a constraint could productively be articulated.

4.2. Speaker's communicative intention

In addition to reusing contextually primed numerals and quantifiers, it is intuitively apparent that the speaker is able to introduce new numbers or quantifiers into the discourse at will, even if they are not present in the observable context. This possibility does not appear to be encompassed by the existing constraints, even if we take the definition of 'primed' numerals (and quantifiers) to include those present in the non-linguistic context. We could address this by positing a constraint requiring the speaker to use a number or quantifier that they wish to make salient in the discourse, which we could characterise as 'speaker's faithfulness to their communicative intention'.

This approach may be logically necessary unless we assume that speakers are behaving entirely deterministically, but it appears to make predictions that are unfalsifiable given our present state of knowledge (we could write off any inexplicable usage as 'faithfulness to the speaker's communicative intention'). For instance, we do not seem to be able to make predictions about the speaker's choice if prompted to "pick a number between one and ten". However, it should be noted that some cases of a speaker introducing a quantity expression reflect that speaker transcoding numerical information that is present in the observable

context, such as by telling the time, reporting a count, or relaying data from a newspaper article. In order to accommodate these cases within our model, we favour the use of a broad definition of context.

4.3. Hearer's knowledge state

For the speaker to make a useful contribution to the communicative exchange, the information they convey must include information that was not previously known to the hearer. Under this assumption, a speaker might be expected to adjust the choice of utterance according to the current knowledge state of the hearer.

This represents a refinement of the notion of informativeness present in the model. For instance, consider a pair of utterances such as (18) and (19).

(18) More than 10 people got married in Cambridge today.

(19) More than 11 people got married in Cambridge today.

According to the constraints as discussed so far, (18) incurs an additional violation of informativeness by comparison with (19). However, a typical hearer might be presumed to understand that the number of people getting married in Cambridge today would be highly likely to be even, and therefore that both (18) and (19) are paraphrasable as "At least 12 people/6 couples…". So, if the hearer's perspective is considered, both (18) and (19) convey the same information.

The problem with incorporating this into the constraint-based model is that the model is speaker-referring. That is to say, it is a model of the speaker's decision procedure, or how the speaker uses the information available to select the optimal utterance. Generally we cannot

presume that the speaker is privy to the hearer's informational state, although once again we might observe that speakers are typically adept at providing information that was not previously known to their interlocutors. Therefore if we were to include this constraint in the model, it would necessarily have to make reference to the speaker's perception of their interlocutor's knowledge state. By analogy with the existing informativeness constraint, it could be articulated as a requirement that the speaker attempt to leave the hearer in as little doubt as possible as to the quantity being conveyed by the utterance.

By comparison to the existing informativeness constraint, this proposed constraint is more naturally 'contextual', inasmuch as it predicts the speaker's choice of utterance will be conditioned by a further aspect of discourse context, namely the (perceived) knowledge state of the interlocutor(s). However, in order to articulate and use this constraint, we would need to make further assumptions as to the speaker's ability to track epistemic state, and attempt to disentangle this from the speaker's own communicative intention (which of course might also vary according to who the hearer is). Given the current state of knowledge in this area, we feel that any attempt to codify such a constraint would be highly speculative, so we leave it aside in what follows.

5. Constraints and a definition of 'relevant context'

So far we have considered how we can use constraints to capture the effect of context on numerical quantifier usage and interpretation. In this section, we examine the reverse perspective, considering instead how the set of constraints suggested naturally constitutes a proposal as to the nature of 'relevant context' for the pragmatics of quantity expressions. We then discuss the extent to which this proposal can be separated from the specific formalism used in the constraint-based model.

5.1. Constraint violations and irrelevant contextual information

In the constraint-based model considered here, the optimal output is selected by evaluating the extent to which candidate outputs violate markedness and faithfulness constraints. There are no *a priori* restrictions on the outputs than can be considered, so in principle the set of candidate outputs is identical and exhaustive in each case. (This infinite set of possible outputs can rapidly be thinned out as all but a small finite number of candidates will incur very numerous constraint violations.) Any given candidate output will incur exactly the same markedness violations every time it is considered (as these only relate to the form of the output itself), and therefore the fact that different utterances are optimal in different situations depends entirely upon the action of faithfulness constraints.

Within such a model, contextual information can influence the speaker's choice of utterance only through faithfulness constraints. Consequently, contextual information that is not relevant to assessing whether or not faithfulness constraints are violated cannot influence the speaker's decision-making process. Meanwhile, the definition of each faithfulness constraint specifies precisely what would constitute a violation. Therefore, we can interpret a proposed set of constraints as a proposed definition of relevant context, simply by considering the union of all the contextual factors that are referred to by markedness constraints in the model.

If such a model is intended to be psychologically realistic as an account of speaker behaviour, it further follows that this model constitutes a proposal as to the contextual state that is tracked by the speaker, as this is necessary for the evaluation of constraint violations. This claim is relatively trivial with respect to the constraints on numeral and quantifier priming, as the prior mention of entities could be presumed to activate them in the mind of discourse participants, without supposing the existence of any additional machinery. However, in the case of granularity, this is not so clear: it might be argued that the existence of a granularity

constraint supposes that the speaker has some means of keeping track of the discourse granularity level.

The model being considered here does not incorporate the constraints referring to the question under discussion or the hearer's knowledge state, discussed in section 4 of this chapter. Therefore, as it stands, the model posits that these constraints are not necessary. It therefore tacitly proposes that aspects of context referred to by these constraints, and not by any others in the system, do not influence the speaker's choice of utterance, and that the speaker does not need to keep track of these considerations.

If a model of this type were to prove satisfactory in its current form, it would follow that the contextual factors referred to by its faithfulness constraints were sufficient to capture the influence of context on the choice of numerically-quantified expression. Hence, this approach presents a possible route towards determining which aspects of context are relevant: given a descriptively adequate model, we could immediately discern that considerations absent from the definition of its constraints were irrelevant. However, it does not follow that all the contextual factors mentioned by the faithfulness constraints of a descriptively adequate model are necessarily relevant to the speaker's decision-making process. Such a model could be over-engineered and contain constraints that are not necessary. Moreover, the adequacy of such a model does not exclude the possibility that an alternative model might be descriptively adequate while encoding a different set of contextual factors.

5.2. Applicability to other formalisms

In the preceding subsection we discussed how a constraint-based model of the type proposed by Cummins (2011) gives rise to a definition of relevant context. However, this model is

couched in terms of classical OT, which might prove unsatisfactory for various reasons. What happens if we consider alternative formalisms?

One interesting possibility is stochastic OT (Boersma 1997), which answers the criticism that the classical OT approach is too restrictive to capture the variability that is characteristic of speaker behaviour. In stochastic OT, an individual speaker's constraint ranking varies to some extent, rather than being fixed in perpetuity. Crucially, none of the discussion in section 5.1 relies upon the constraint ranking being fixed. Under the assumptions of stochastic OT, it is still impossible for contextual factors to influence the choice of utterance other than via faithfulness constraints, and so the specification of those faithfulness constraints would still encode a proposal as to the nature of relevant context. Again, if a stochastic OT account were to prove descriptively adequate, this would argue forcibly against the relevance to the speaker of contextual factors that were not referred to by the faithfulness constraints proposed within that particular account.

More generally it might be argued that the whole OT approach is potentially unsatisfactory as an account of speaker behaviour. The proposed constraints are individually functionally motivated, and the results in section 3 support the contention that they interact in determining the speaker's optimal utterance. However, it is possible that this interaction is evaluated by the speaker in an entirely different way. For instance, each candidate utterance might be given a score according to how well it satisfies each constraint, with the 'winning' utterance being that with the highest total score. From this perspective, the proposal could be seen as something akin to an attempt to identify the individual constituents of 'relevance', albeit with greater focus on speaker effort than is customary in relevance-theoretic accounts (although Wilson and Sperber (2002: 257) note explicitly that considerations of the speaker's preferences influence the choice of utterance). However, even in this setting, the above argument goes through. In this case, contextual factors that are not referred to by the

constraints do not have any effect on determining the 'final score' for any candidate utterance. Once again, the specification of the constraints amounts to a hypothesis about the nature of relevant context.

Still another analytic possibility would be to model the system as a connectionist network in which the input layer represents the situation and the output layer represents the utterance. The faithfulness constraints of this model could then be identified with the connections between these layers. Given an input, the optimal output could be selected by Harmony Maximisation (Smolensky 1986). In this model we could identify 'relevant context' as that which is encoded by those nodes on the input layer that are connected to nodes on the output layer with non-zero weights. Once again, contextual material that was not referred to by the faithfulness constraints – i.e. that which was represented by nodes on the input layer which did not have connections to nodes on the output layer – would be irrelevant to the process of determining the optimal output.

5.3. Interim summary

In the above sections, we discuss how a set of faithfulness constraints can be read as a proposal as to the nature of relevant context. We then show that, although the terminology of 'faithfulness constraints' is specific to OT, this perspective on context is also available within alternative formalisms. In section 7 we state this generalisation and consider its potential usefulness. However, before doing so, we turn briefly to the question of how the hearer is able to use context to interpret the speaker's utterances within this approach, and how additional information about context is naturally conveyed by the speaker.

6. Context furnished by the hearer

In discussing the hearer's interpretation of numerically-quantified statements under this model (e.g. sections 3.1 and 3.2), we have assumed that the hearer is privy to the same contextual information as the hearer. For instance, in (20), A is aware that the numeral 60 is contextually activated and adjusts the interpretation of B's utterance accordingly.

(20) A. This display case holds 60 CDs. How many CDs do you have?

B. I have more than 60 CDs.

The role of context in interpretation within this model is diametrically opposite to its role in production. The speaker is presumed to encode his intention optimally, taking into account the constraints (which include reference to context). The hearer is then presumed to decode that intention by taking the utterance that results and allowing for the constraints[6]. The constraints, from this point of view, behave like the key to a cipher: the speaker's utterance reflects both their intention and the constraints by which they are bound, but the hearer is already privy to the constraints and is only interested in recovering the speaker's intention.

Nevertheless, this view still supposes that the speaker's utterance conveys information about the constraints as well as about the speaker's intention. With respect to considerations of markedness, we might expect fairly reliable agreement between two competent speakers – it is perhaps relatively uncontroversial what constitutes a prolix expression, a complex quantifier or a round number – but with respect to contextual factors it is possible for the two interlocutors to become de-aligned.

Two consequences would be expected to arise from this, both of which seem intuitively plausible in real-life interactions. One is that miscommunication might arise. As an artificial

[6] This is dissimilar to a bidirectional OT model in that there is no one-one mapping between form and meaning, so the hearer's task is fundamentally different from the speaker's task. The hearer may in principle recover a different intention to that encoded by the speaker, as we discuss below.

example, consider (21), said of a player in a tennis match. A hearer who is unaware of the granularity applicable to tennis scores (0, 15, 30, 40, game) might be entitled to conclude that the speaker was not in a position to say (22), whereas according to numerical considerations this is in fact entailed by (21).

(21) He's scored more than 15 in every game.

(22) He's scored more than 20 in every game.

A potentially more interesting consequence is that the hearer may be able to draw inferences about the context based on the speaker's utterance. Consider (6), repeated below as (23).

(23) Will Tiger win more than 18 majors?

Earlier we argued that the use of "more than 18" is licensed only because 18 is a contextually salient number. A hearer, confronted with this sentence, is entitled to reason the same way, and should arrive at the conclusion that 18 is a salient number (and might, given certain background information, also correctly infer that it is the record for the most majors won by a single golfer). Thus the speaker's utterance conveys something more than the semantic content of the utterance: it also conveys the especial relevance of the number being referred to. This can be exploited by the speaker to convey additional information efficiently. Given that the non-round numeral does not give rise to scalar implicatures, we might see this as the speaker exploiting the intrinsic pragmatic uselessness of the utterance to good pragmatic effect.

The results of Cummins, Sauerland and Solt (submitted) also tend to support this observation. As discussed earlier, there are anomalies in the interpretation of modified fine-grained numerals (such as "more than 93"). Developing the discussion in section 3.2 of this chapter, we might suppose a hearer reasons as follows: if a speaker utters "more than 93", *either* they

do not have better knowledge *or* there is some particular reason why 93 was mentioned. In either case, the utterance should give rise to a weaker implicature than would otherwise theoretically be predicted.

If this analysis is approximately correct, it suggests that the hearer is at once pragmatically aware and anxious to draw contextual enrichments, and ready to acknowledge the limitations of their own knowledge. The hearer is generally not privy to the mental state of the speaker, and consequently cannot know for certain what is and is not 'primed' from the speaker's perspective, although they may be confident that a number just mentioned in the speaker's presence should qualify as 'primed'. This further suggests that, in general, the hearer should draw pragmatic enrichments with caution – even if the utterance is "more than 200", this might reflect contextual activation of that particular numeral, which might moderate the implicatures that are yielded. However, the idea that the hearer should exercise caution in implicature is a very general one: even in canonical cases such as (24), it is possible that B's response reflects the strength of his feeling about the proposition he expresses rather than the falsity of an informationally stronger proposition.

(24) A. Did you meet Jane's parents?

B. I met her father.

Thus it appears that the speaker is able to convey information about the context as well as the underlying intention. The nature of the context-referring constraints delimits the extent to which the speaker is able to exploit this channel. The numeral priming constraint seems readily to be exploited to convey the relevance of the numeral used. Similarly, a speaker can convey that they consider a particular level of granularity appropriate (e.g. by saying "it's 6:04"), although the pragmatic consequences of that appear limited. Of course, the model does predict that speakers can use utterances of this type to prompt for a particular type of

response from their interlocutor (e.g. a precise one), and this also involves additional pragmatic information being conveyed by exploitation of faithfulness constraints.

7. Discussion and conclusions

In this chapter we explored how a specific account of quantifier usage gives rise to a proposal as to the nature of relevant context. We briefly examined the constraints proposed by Cummins (2011) as an example of how faithfulness constraints can be construed as a claim about which aspects of context influence utterance selection. We then considered the application of this idea to more general models of speaker behaviour, and briefly explored how the speaker can exploit constraints to convey information about context.

The general premise we are adopting is that utterances (in this case, numerically-quantified expressions) are selected according to how well they meet two sets of criteria, one concerned with their intrinsic structure and one concerned with their relation to their context. The general conclusions we draw are that the specification of this latter set of criteria constitutes a specific proposal about the nature of relevant context, and that a descriptively adequate set of criteria must refer to all relevant contextual factors. That is to say, an account will not be descriptively adequate if it omits reference to any relevant aspect of context, and therefore an account that is descriptively adequate will necessarily encompass all such aspects of context.

The value of this viewpoint to the process of modelling context depends to a large extent on whether a model of the type presented here (or any of the variants discussed in section 5.2) can approach descriptive adequacy as an account of speaker behaviour. However, there are advantages to an analysis of this type, not least that it mandates a precise specification of contextual factors whenever these are appealed to, and presents a framework within which

predictions about their impact on the speaker can be quantified. It forces us to dissect the notion of context and treat its component parts in a suitably precise way. In this way, the approach presented here may enable us to tackle in relatively easy stages the huge task of ascertaining the nature of relevant context.

References

Barwise, J. and Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy,* 4: 159-219.

Blutner, R. (2006). Embedded implicatures and optimality theoretic pragmatics. In T. Solstad, A. Grønn and D. Haug (eds.), *A Festschrift for Kjell Johan Sæbø: in partial fulfilment of the requirements for the celebration of his 50th birthday*. Oslo.

Boersma, P. (1997). How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences*, 21: 43-58.

Cummins, C. (2011). The interpretation and use of numerically-quantified expressions. Unpublished PhD thesis, University of Cambridge.

Cummins, C. and Katsos, N. (2010). Comparative and superlative quantifiers: pragmatic effects of comparison type. *Journal of Semantics*, 27: 271-305.

Cummins, C., Sauerland, U. and Solt, S. (submitted). Granularity and scalar implicature in numerical expressions.

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

Geurts, B., Katsos, N., Cummins, C., Moons, J. and Noordman, L. (2010). Scalar quantifiers: logic, acquisition, and processing. *Language and Cognitive Processes*, 25: 130-48.

Geurts, B. and Nouwen, R. (2007). 'At least' et al.: the semantics of scalar modifiers. *Language*, 83: 533-59.

Grice, H. P. (1975). Logic and Conversation. In Cole, P. and Morgan, J. L. (eds.), *Syntax and Semantics*, Vol. 3. New York: Academic Press. 41-58.

Gualmini, A. (2004). Some knowledge children don't lack. *Linguistics*, 41: 957-82.

Hendriks, P. and de Hoop, H. (2001). Optimality Theoretic semantics. *Linguistics and Philosophy*, 24: 1-32.

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*. Berlin: Mouton de Gruyter. 439-58.

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*. Stanford: CSLI Publications. 109-132.

McCarthy, J. J. (2002). *A Thematic Guide to Optimality Theory*. Cambridge: CUP.

Musolino, J., Crain, S. and Thornton, R. (2000). Navigating negative quantificational space. *Linguistics*, 38: 1–32.

Pickering, M. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27: 169-226.

Prince, A. and Smolensky, P. (1993). *Optimality Theory: Constraint Interaction in Generative Grammar*. Rutgers University Center for Cognitive Science Technical Report 2.

Roberts, C. (1996). Information structure in discourse: towards an integrated formal theory of pragmatics. In Yoon, J.-H. and Kathol, A. (eds.), *OSUWPL Volume 49: Papers in Semantics*. Columbus, OH: Ohio State University Department of Linguistics.

Smolensky, P. (1986). Information processing in dynamical systems: foundations of harmony theory. In Rumelhart, D. E., McClelland, J. L. and the PDP Research Group, *Parallel Distributed Processing: Explorations in the microstructure of cognition. Volume 1: Foundations*. Cambridge, MA: MIT Press/Bradford Books. 194-281.

Van der Henst, J. B., Carles, L. and Sperber, D. (2002). Truthfulness and relevance in telling the time. *Mind and Language*, 17: 457-66.

Van der Henst, J. B. and Sperber, D. (2004). Testing the cognitive and communicative principles of relevance. In Noveck, I. and Sperber, D. (eds.), *Experimental Pragmatics*. Basingstoke: Palgrave Macmillan. 141-69.

Wilson, D. and Sperber, D. (2002). Truthfulness and relevance. *Mind*, 111: 583-632.

Zondervan, A. (2007). Effects of Question Under Discussion and focus on scalar implicatures. In Kluck, M. E. and Smits, E. J. (eds.), *Proceedings of the Fifth Semantics in the Netherlands Day (SiN V)*. 39-52.