Generation of Quantified Referring Expressions: Evidence from Experimental Data

Dale Barr
Dept. of Psychology
University of Glasgow
dale.barr@glasgow.ac.uk

Kees van Deemter
Computing Science Dept.
University of Aberdeen
k.vdeemter@abdn.ac.uk

Raquel Fernández
ILLC
University of Amsterdam
raquel.fernandez@uva.nl

Abstract

We present the results from an elicitation experiment in which human speakers were asked to produced quantified referring expressions (QREs), as in *'The crate with 10 apples'*, *'The crate with many apples'*, etc. These results suggest that some subtle contextual factors govern the choice between different types of QREs, and that numerals are highly preferred for subitizable quantities despite the availability of coarser-grained expressions.

1 Introduction

Speakers can express quantities in different ways. For instance, a speaker may specify a meeting time with the expression *'in the morning'* or with the more precise, numeric expression *'at 10:30am"*; she may choose to specify a temperature as *'5 degrees Celsius'* or instead use the less precise but more qualifying expression *'cold'*. One area of NLG where these choices are important is the generation of referring expressions. In particular, a referent may be identified by means of some quantitative value or other (e.g., *'the tall man*; *'the man who is 198cm tall'*), or by means of the number of other entities to which it is related. Henceforth, let's call these *quantified* referring expressions (QREs). An example of a QRE arises, for instance, when a person is identified by means of the number of his children (*'the man with 5 daughters'*), when a directory is identified by means of the number of files in it (*'the directory with 520/many PDF files in it'*), or when a crate is identified by means of the number of apples in it (*'the crate with 7 /a few apples'*).

Green and van Deemter (2011) asked under what circumstances it might be beneficial, for a reader or hearer, for referring expressions of this kind to contain vague expressions (e.g., like *many*). The present paper addresses the same phenomena focussing, more broadly, on all the different ways in which reference may be achieved; unlike these previous authors, we shall address this question from the point of view of the speaker, asking how human speakers refer in such cases, rather than how useful a given referring expression is to a hearer (e.g., as measured by their response times in a manipulation task).

We start by making our research questions more precise in the next section. We then describe the production experiment we run online in Section 3 and present an analysis of the data in Section 4. We end with some pointers on how our results could inform an NLG module for QREs.

2 Research Questions

Suppose you want to point out one crate amongst several crates with different numbers of apples. You may use a numeral (*'the crate with seven apples'*) or, if the crate in question is the one with the largest or smallest amount of apples, you may use superlatives (*'the crate with the most apples'*), comparatives (*'with more apples'*) or vague quantifiers (*'with many apples'*); if your crate is the only one with any apples in it at all, you might simply say *'the crate with apples'*). In many situations, several of these options are applicable. It is not obvious, however, which of these is preferred. The Gricean Maxim of Quantity (Grice, 1975) urges speakers to make their contribution as informative as, but not more informative than, it is required for the current purposes of the exchange. This might be taken to predict that speakers will tend to use the most coarsely grained expression that identifies the referent (unless they want some nontrivial implicatures to be inferred). This would predict, for example that it is odd to say *'the box with 27 apples'* when *'the box with apples'* suffices, because the latter contains a boolean property (contains apples), whereas the former relies

Proceedings of the 14th European Workshop on Natural Language Generation, pages 157–161,
Sofia, Bulgaria, August 8-9 2013.

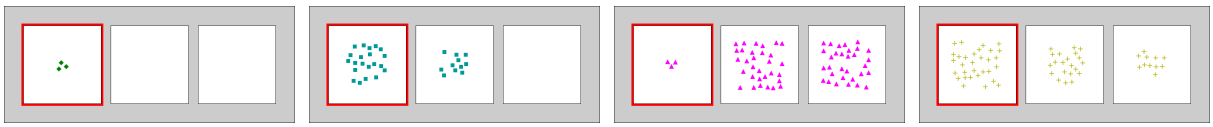

Figure 1: Sample stimuli in contexts X__, XY_, XYY with big gap, and XYZ with small gap.

on a special case on what is essentially much more finely grained property (contains x apples).

Our hunch, however, was that this is not the whole story. For example, the literature on human number processing suggests that numbers below 5 or 6 are handled almost effortlessly; these numbers are called *subitizable* (Kaufman et al., 1949) Furthermore, we hypothesized that it matters to what extent the number of apples in the target crate "stands out". We had the following expectations:

1. Speakers do not always use the coarsest-grained level that is sufficient.
2. Whether a quantity is subitizable or not interferes with the speakers' choice.
3. The frequency of vague forms (such as *'many'*) will be higher in contexts where the gap between the target quantity and the quantities in the distractors is large than when it is small.[1]

We wanted to put these ideas to the test and, more generally, find out how human speakers use QREs in different contexts. Our interest was also in creating a corpus of human-produced QREs that can serve future research.

3 Experimental Setup

The elicitation experiment was run online. Subjects first encountered a screen with instructions. They were told that they would be presented with situations consisting of three squares, with each of them having none, one or more shapes in it. In each of these situations, one of the three squares would be highlighted and subjects were asked to describe this target square in a way that would enable a reader of their expression to identify it. Subjects were told that the recipient of their description may see the three squares arranged differently on the screen with their contents possibly being scrambled around. That is, they were indirectly asked to concentrate on the *quantity* of shapes in the squares (rather than on their relative position or on the spatial configuration of the shapes in them). Figure 1 shows some sample stimuli.

The experiment included a total of 20 items, generated according to the following parameters:

- *Subitizability*: the amount of shapes in the target is within the subitizable range (SR) (1-4 shapes) or within a non-subitizable range (NR); we included three non-subitizable ranges, with around 10, 20, and 30 shapes, respectively.
- *Context*: we considered four types of scenarios:
  1. X__: only the target square is filled.
  2. XY_: two squares are filled.
  3. XYY: all squares filled; with two ranges.
  4. XYZ: all squares filled; with three ranges.

  The symbol X in the first position stands for the referent square, while the symbols in the other two positions indicate for each of the other two squares whether it contains a number of shapes within the same range as the referent square (X), within a different range (Y/Z), or whether it does not contain any shapes at all (_).
- *Relative Size*: the target contains either the smallest or the largest amount of shapes.
- *Gap Size*: there is either a big or a small quantity difference between the target and other squares. A big gap size is only possible with target squares that contain the largest amount of shapes within a non-subitizable range and those that contain the smallest amount of shapes within a subitizable range.

Participants were recruited by publishing a call in the Linguist List. A total of 82 subjects participated in the experiment, including participants who only responded to some items. We eliminated 6 sessions where the participant had responded to less than 10 items. The final dataset includes 76 participants and a total of 1508 descriptions.

4 Results

Each description produced by the participants was annotated with one of the categories in Table 1.

[1] Later on we refer to vague forms as "base", a common term used to describe the vague, unmodified form of relative scalar adjectives (e.g., *tall*) as opposed to their comparative (*taller*) and superlative (*tallest*) forms.

Category	Examples
ABS [absolute]	*the one with pacmans* / *the square that's not blank*
BASE [base]	*the square with lots of dark dashes* / *it has a few crosses in it*
COMP [comparative]	*the one with fewer dashes* / *the square with more crosses in it*
NUM [numeric]	*the square with 11 black dots* / *3 grey ovals*
SUP [superlative]	*it has the largest number of purple squares* / *the square with the least minuses*
OTH [other]	*about a dozen blue diamonds* / *big droup of circles in the centre*

Table 1: Categories used to code the expressions produced by the participants.

The classification was first done automatically by pattern matching and then revised manually.

To analyse the data, we used mixed-effects logistic regression with crossed random effects for subjects and items (Baayen et al., 2008). All models had by-subject and by-item random intercepts, and by-subject random slopes for the within-subject factors of context and range (subitizability). The models were fit using maximum likelihood estimation with p-values derived from likelihood ratio tests. Model estimation was performed using the lme4 package (Bates et al., 2013) of R statistical software (R Core Team, 2013).

Table 2 shows the overall distribution of expression types used by the participants. As can be seen, numerical expressions were the most common type of expression used overall (65%). We found, however, that there was a strong subitizability effect in the use of these expressions: for non-subitizable targets, subjects used numerical expressions only 39% of the time, while for subitizable targets they did so 90% of the time. This main effect of subitizability was significant ($\chi^2(1) = 47.92$, $p < .001$). There was high variability across subjects in the effect ($\chi^2(1) = 25.00$, $p < .001$), with a higher rate of numerical expressions associated with a smaller effect of subitizability ($r = -.61$). Note that 17 of the 82 subjects ($\sim 20\%$) *always* used numerical expressions, even when the target was not subitizable. Of the remaining 65 subjects, 64 show a very significant preference for using numeric expressions to describe targets within the subitizable range.

Figure 2 shows the proportion of expression types for each type of context and subitizability condition.[2] Sensitivity to context differed for subitizable and non-subitizable targets, supported by a reliable interaction between these factors ($\chi^2(1) = 17.31$, $p < .001$). Despite the strong overall preference for numerical expressions with subitizable targets, the effect of context was still reliable ($\chi^2(1) = 22.63$, $p < .001$). For subitizable targets (Figure 2, bottom row), numeric expressions were almost always used (96%) except in contexts where the target was the only filled square (X__). In this context, participants occasionally used absolute expressions instead (e.g. *the one with shapes*) 33% of the time. In sum, subitizable targets overwhelmingly triggered the use of numerals, predominating even when a Gricean account would prefer coarser-grained expressions.

	ABS	BASE	COMP	NUM	SUP	OTH	Total
NR	73	33	26	294	308	17	751
SR	51	1	0	684	21	0	757
Total	124	34	26	978	329	17	1508

Table 2: Row counts of expression types for non-subitizable (NR) and subitizable (SR) targets.

For non-subitizable targets (first row of plots in Figure 2), in contexts without distractors (X__) absolute expressions were preferred over numerical ones; this differed from the behaviour of subitizable targets in this context, where numerical expressions predominated ($\chi^2(1) = 4.25$, $p = .039$). In contexts with non-empty distractors (XY_, XYY, and XYZ), expressions other than numeric are used significantly more often than they were for subitizable targets ($\chi^2(1) = 52.93$, $p < .001$). Superlative expressions (e.g. *the square with the least dots*) were preferred in contexts where the three squares were filled ($\chi^2(1) = 7.74$, $p = .005$). In contexts with one distractor (XY_), superlatives were also rather common, and comparative expressions (e.g. *the one with fewer dashes*) occurred at higher rates than in other types of context ($\chi^2(1) = 42.34$, $p < .001$).

The comparison between the contexts with two distractors (XYY and XYZ) suggests that they differed largely in the use of vague expressions (BASE; e.g. *the one with many diamonds*), which had a higher rate in context XYY where there were only two quantity ranges ($\chi^2(1) = 5.01$,

[2] Category OTH (other) is not shown in Figure 2 to avoid clutter. Table 2 shows the row counts for all categories.

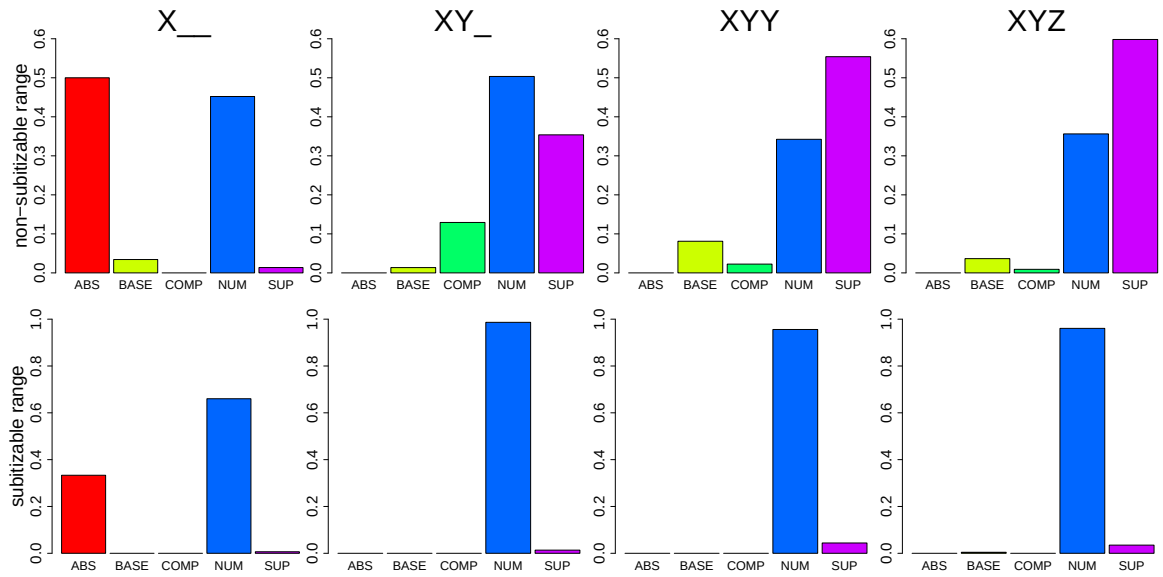

Figure 2: Proportion of expression types in each context for subitizable and non-subitizable targets.

$p = .025$). For this context we also found an effect of gap size (see Figure 3): the relative odds of choosing a vague expression over a numeric or superlative one is significantly higher when there is a big difference between the target quantity and the distractor quantities ($\chi^2(1) = 5.68, p = .017$); that is, when the chance of there being borderline cases is reduced. A small gap between the quantities makes the preference for superlative (and thus non-vague) expressions stronger.

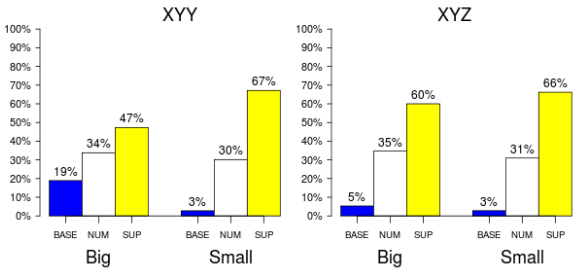

Figure 3: The effect of gap size.

5 Conclusions

In line with our expectations (see Section 2), our data are not easy to reconcile with the type of Gricean account that predicts a preference for the most coarsely grained QRE that identifies the target. The most obvious deviation from this Gricean account arises from the subitizable items in our study, where numerical expressions turned out to be much preferred over other QREs. The natural explanation seems to be that such expressions *come naturally* to speakers (and to hearers too as shown by Green and van Deemter (2011)). In other words, our study suggests an intriguing variant on Grice, in which the most relevant factor is not one of *informativeness* – as Grice's writings suggest – but one of effort. It suggests that speakers tend to produce expressions that identify the referent *with least effort*.

Our expectation 3 was also confirmed: vague forms (`BASE`) are more frequent with big gap sizes, although they are not produced with high frequency. (The same pattern of results was found by van Deemter (2004)). Thus, in the scenarios we considered vague QREs are never the most favoured option. The high frequency of superlatives over comparatives is also noteworthy. Comparatives are used very seldom overall but are more frequent in contexts with only one distractor (`XY_`). This indicates that some speakers opt for a less strong expression than a superlative (an expression that means *more than x* rather than *more than any other x*) in contexts where this does not lead to ambiguity. However, numerals and superlatives are still largely preferred in those contexts.

These observations suggest that a given type of situation (i.e., a given context + subitizability condition) should not always map to the same type of QRE. If human QRE behaviour is to be mimicked, the best approach seems to be to use a stochastic NLG program that seeks to replicate the frequencies that are found in human usage.

The collected data is freely available at `http://www.illc.uva.nl/~raquel/xprag/`.

References

R. Baayen, D. Davidson, and D. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412.

D. Bates, M. Maechler, and B. Bolker, 2013. *lme4: Linear mixed-effects models using S4 classes*. R v. 0.999999-2.

M. Green and K. van Deemter. 2011. Vagueness as cost reduction: An empirical test. In *Proc. of Production of Referring Expressions workshop at CogSci 2011*.

H. P. Grice. 1975. Logic and conversation. In *The Logic of Grammar*, pages 64–75. Dickenson.

E. Kaufman, M. Lord, T. Reese, and J. Volkmann. 1949. The discrimination of visual number. *American Journal of Psychology*, 62(4):498–525.

R Core Team, 2013. *R: A Language and Environment for Statistical Computing*. R Foundation. v. 3.0.0.

K. van Deemter. 2004. Finetuning NLG through experiments with human subjects: the case of vague descriptions. In *Proc. of the 3rd INLG Conference*.