# Is Vagueness Beneficial for Hearers? Evidence from Experiments

Matt Green
Computing Science
University of Aberdeen

Kees van Deemter
Computing Science
University of Aberdeen

Abstract

Much of everyday language is vague, but the causes of this phenomenon are not well understood. Consequently, it is difficult for the designers of a Natural Language Generation (NLG) system to know when and how to let the system generate vague expressions. The present paper is an attempt to find out what benefits vagueness might have for readers.The paper reports on a series of experiments that aim to separate the utility of vagueness (as defined by the existence of borderline cases) from the utility of other factors that tend to co-occur with vagueness. We argue that the evidence so far supports a view of vagueness where the benefits that vague terms exert are due to other influences, rather than to vagueness itself. These factors include: low granularity; the use of evaluative words; the avoidance of overtly numerical words; the existence of comparison strategies; and, lastly and more tentatively, a phenomenon that we call range reduction. Although it is possible that other types of vague expressions (i.e., vague words outside referential noun phrases) behave differently, our work suggests that vagueness itself may not increase the utility of an expression. The paper concludes with a brief discussion of the implications of our work for practical NLG.

## Introduction

Vagueness pervades the language that we use on a daily basis, and the challenge of understanding vague language has been a prominent concern in many areas of logic and linguistics, involving both theoretical and applied work, including the area known as Natural Language Generation (NLG).

NLG systems transform data and formulas into language (e.g., Reiter & Dale, 2000). NLG systems routinely make decisions between different formulations of the same information. For example, if the temperature is 27.2 degrees Celsius, this could be expressed as "27.2 degrees", "approximately 27 degrees', "above 25 degrees", or "warm", and the system must decide between these. The problem is especially important for NLG systems that take numbers as input, for example in the generation of textual weather reports from numerical weather data such as temperature and wind speed (Goldberg, Driedger, & Kittredge, 1994; Turner, Sripada, Reiter, & Davy, 2006), and medical decision support on the basis of clinical measurement such as oxygen saturation, heart rhythm, etc. (Hripcsak, Elhadad, Chen, Zhou, & Morrison, 2009; Hunter et al., 2008; Portet et al., 2009). Such systems are often forced to make decisions concerning the level of precision in the utterances that they generate on the basis of little more than intuition. Even when NLG systems are designed to mimic human language use (e.g., Konstas and Lapata 2013), there is no guarantee that these decisions taken by these systems benefit readers. A better understanding of the benefits (for readers and hearers) of different precision levels would allow these systems to become more useful. The present article investigates the benefits, or de-benefits, of vagueness.

Language use may be called vague for various reasons.[1] In most academic use though, the word 'vagueness' has a specific meaning. Keefe and Smith, for example, state "vague predicates have borderline cases, have fuzzy boundaries, and are susceptible to sorites paradoxes" (Keefe & Smith, 1996, p. 4), also Egre and Klinedinst (2011)). The crucial criterion is the existence of borderline cases: "a word is precise if it describes a well-defined set of objects. By contrast, a word is vague if it is not precise" (Lipman, 2009, p. 1). A typical example is the word "tall", as applied to people for example, because here is no precise, known height which separates those who are tall from those who are not. The crucial point is that "tall" admits borderline cases (i.e., people who may or may not count as tall), which are the hallmark of vagueness as we use the term.

Linguists, philosophers of language, and more recently game theorists, have asked why natural languages contain so many vague expressions (Lipman, 2000, 2009). By introducing borderline cases, these expressions create potential misunderstandings, thereby creating "a worldwide several-thousand year efficiency loss" (Lipman, 2009, p. 1). Lipman explains the point by means of a scenario in which a speaker describes a person to a hearer, who needs to identify that person in the arrivals hall of of an airport. In such a scenario, a precise description of the person's height (e.g., "The person's height is 187.96 cm") would be more useful than a vague one ("The person is tall"). Lipman uses this scenario to explain why standard game theory models of communication (e.g., Crawford & Sobel, 1982) predict that, under certain conditions, a crisp act of communication will always have more utility than a vague act that communicates the same state of affairs.

Lipman argued that the efficiency loss resulting from vague expressions would be unlikely to have arisen unless there are advantages as well as disadvantages associated with vague expressions. Lipman asked, essentially, what these advantages might be. Several tentative answers to Lipman's question have been offered (see van Deemter, 2009, 2010). Prominent among these answers is the idea that vague expressions are easier to process, by a speaker and/or a hearer, than expressions that are not vague (i.e., crisp) (e.g., Lipman,

[1] See e.g. the entry "vague" in (Allen, 2000).

2009; De Jaegher, 2003; van Rooij, 2003). For example, Lipman (2009, p. 11) writes: "For the listener, information which is too specific may require more effort to analyze". We shall refer to this as the *cost reduction* hypothesis.

Charting the utility of vagueness is the attested aim of a small number of studies, but most of these have focussed on vagueness in a different sense, and focussing on different types of benefits for hearers. Two recent studies can illustrate both issues.

In a series of studies of behaviour modification, Mishra, Mishra, and Shiv (2011) manipulated the presentation format of information about quantities in the domains of mental acuity, physical strength, and weight loss. In the weight loss study, participants were told that the study was designed to test the validity of a new (actually fictitious) health index, the HHI (Holistic Health Index). They were told that an ideal HHI score lies in the range of 45 to 55. In a longitudinal study, participants submitted their height, weight, hydration level, gender, and age to a computer each week. Participants were told that two algorithms would be used to compute their HHI, and that it was possible that the two algorithms might give different values initially, but would converge over the course of the study to a single value. They were also told that if the two algorithms did give different values, then the true score lay between the two values. In one condition, which the authors called the precise condition, the two algorithms gave the same score. In the other condition, which the authors called the vague condition, one algorithm added 3% to the score while the other algorithm subtracted 3% from the score, yielding a range of values whose midpoint was the same as the two values given in the precise condition.

One group of participants was given HHI scores in the ideal range: for this group their weight loss did not differ depending on whether they were given vague or precise HHI values. However for the other group, who were given HHI scores outside the ideal range, their weight loss was significantly greater if they were given vague HHI scores than if they were given precise HHI scores. The authors explain the improvement in the vague condition for this group as resulting from the participants' freedom to think of themselves a positioned on one end of the range - the end closest to the ideal HHI scores. This "illusion of proximity" (Mishra et al., 2011, p. 4) to the goal is argued to allow participants to generate positive expectancies that lead to behaviours that improve performance. In contrast, in the precise conditions, participants did not have this freedom of interpretation, and could not distort the information to bring about the beneficial *illusion of proximity*. These results are interesting, and of obvious potential practical importance. We note, however, that information presented as an exact range of values does not conform with the standard definition of vagueness (Keefe & Smith, 1996; Egre & Klinedinst, 2011), since an exact range does not admit borderline cases. In the terminology of Hobbs (1985), the difference between a range and a single midpoint value is a difference of *granularity*. Furthermore, the experiments of Mishra et al. (2011) did not explore benefits in terms of processing cost, but in terms of long-term behaviour change.

Similar issues arise from the work of Peters et al. (2009). The authors carried out a series of studies where participants were required to rate hospitals based on various sources of information about quality of care. There was a between-subjects manipulation based on numeracy. The format of the information was manipulated within subjects: either numbers only were presented, or both numbers and evaluative categories were presented (e.g., *Poor*, *Fair*, *Good*, *Excellent*, with crisp visual boundary lines between the categories). Results

showed that, for low-numeracy participants, the presence of evaluative categories resulted in a diminished influence of an irrelevant affective state on the ratings. For all participants, the presence of evaluative categories resulted in better decisions and in a greater use of the most important and reliable types of information, such as survival rates.

It is, however, questionable whether the “evaluative categories" manipulation in this study can be considered a manipulation of vagueness. Certainly, terms like *Fair* admit the possibility of borderline cases. However, given that the boundaries between the categories were marked crisply, and that therefore the categories mapped crisply to numerical values, it becomes doubtful whether any borderline cases could be conceived to arise in fact. For example, *Fair* was mapped to 60% – 70% for the variable *percentage of heart attack patients given recommended treatment (ACE inhibitor)*. Accordingly, rather than the vagueness of categories such as *Poor*, Peters et al. emphasise the evaluative content inherent in these categories, and the affective potential of the evaluative content rather than the vagueness of the terms like *Fair*.

**General Methodology.** The experiments reported in the present paper put the cost reduction hypothesis to the test. The question that we are trying to answer is whether vague expressions are processed more easily by readers than crisp ones. Like Lipman, we focus on situations where numerical information is used in order to identify a referent. Reference, in other words, will be the linguistic task on which we focus, partly because of the interest that this topic has recently drawn from the NLG community. In focussing on benefits for the hearer, we will leave aside the question of audience design, leaving this for later research.

In using references to quantities to test the cost reduction hypothesis we are only testing one aspect of vagueness in a particular context. This limits the applicability of our results. However, it has the advantage that it enables us to explore the costs and benefits of vagueness more thoroughly. Since one prevalent view of vagueness is that a vague expression is never preferable to a crisp equivalent, a demonstration of a benefit for vagueness in any context would advance the discussion.

In our experiments we used a speeded forced choice task to compare the processing costs of different references to quantities. In this context, speed and accuracy of responses are the key dimensions on which the different references can be compared. Each stimulus in the experiments was a set of dot arrays containing various number of dots, together with a preceding instruction (in the form of a referring expression) to choose one of the arrays with respect to its cardinality. The participant was asked to respond as quickly as possible while avoiding errors. We manipulated the instructions and the arrays in several ways across a series of four experiments.

There is evidence that when the distance grows between two numbers, they become more easily distinguishable from each other: the *numerical distance effect*, which has been shown for comparing the cardinality of two sets of dots (van Oeffelen & Vos, 1982) and for processing Arabic numerals and number words (Dehaene, 1996). We manipulated the number of dots in each array such that some sets of arrays had smaller numerical distances and others had larger numerical distances. Where a number was mentioned in the instructions, it was always in the form of an Arabic numeral.

There is evidence that when two numbers are presented with the smaller on the left, this left-side presentation facilitates responses indicating the smaller number: the *Spatial-*

*Numerical Association of Response Codes (SNARC)* effect (Dehaene, Bossini, & Giraux, 1993; Gevers, Lammertyn, Notebaert, Verguts, & Fias, 2006). We controlled which side the smaller number appeared on to avoid systematic influences of this effect.

All the experiments shared the following properties: Stimuli were created using the language GNU Octave (Eaton, 2002) and the Psychophysics Toolbox extensions (Brainard, 1997; Kleiner et al., 2007). The position of the dots in each array was randomised per-trial. The order in which trials were presented was randomised per-participant. There were 256 trials, presented in 4 blocks of 64 trials each, between which the participant could rest. A MacBook Pro laptop computer with a 13 inch screen presented the stimuli to the participants and recorded responses. Participants were recruited using email lists at the University of Aberdeen, and paid ten pounds for participating. All participants self-reported fluency in English, and had normal, or corrected-to-normal vision. The experiment was conducted in a quiet room. Participants were asked to respond as quickly as possible while avoiding errors. There was a block of practice trials after which participants could ask any questions, following which the experimenter left the room.

## Experiment One

### Introduction

We used a forced choice task to elicit responses to crisp or vague forms of instructions that required the participant to choose one of two dot arrays by referring to its cardinality. The participant was presented with an instruction in the form *Choose the square with ... dots*. Then a set of two two dot arrays was presented, each in the form of a square containing some number of dots. The participant was required to identify the array that corresponded with the instruction, by pressing the appropriate key, as quickly as possible while avoiding errors. Response time and accuracy were recorded for analysis.

We manipulated how discriminable the dot arrays were by varying the numerical distance between them. One array always contained 25 dots: the other contained either 5, 10, 15, 20, 30, 35, 40, or 45 dots. This gave us numerical distances of 5, 10, 15, and 20, with smaller numerical distances resulting in less discriminable arrays and larger distances resulting in more discriminable arrays.

Our main manipulation was of the vagueness of the instruction, with two levels, *crisp* and *vague*. Assuming the dot array [5, 25], and the instruction referring to the smaller cardinality, the *crisp* instruction was *Choose the square with 5 dots* and its *vague* counterpart was *Choose the square with few dots*.

### Hypotheses

(1) A main effect advantage for vagueness: vague instructions impose a lower cognitive load for the comprehender than crisp alternatives.

(2) A main effect advantage of increasing numerical distance: the task will become easier as the numerical distance increases, because the two arrays are then more discriminable.

(3) An interaction between vagueness and numerical distance: i.e., any facilitation for vagueness should be greater at smaller numerical distances than at larger numerical distances.

## Method

Twenty participants were recruited, aged between 18 and 45 with a median age of 26. On each trial a participant was presented with an instruction to choose one of two dot arrays on screen by reference to its cardinality. Following a keypress to indicate that the participant had read the instruction there was a central fixation cross for 1000 ms, and a blank screen for 500 ms, followed by the array (without repetition of the referring expression). The arrays would stay on screen until the participant responded (there was no timing-out). Response time was measured as the latency between the presentation of the arrays, and the keypress identifying the choice: in this way, the response time was separated from time spent reading the instructions, which is important since we are only interested in the former. A response was counted as erroneous if the square with the wrong number of dots was chosen (when the instruction contained a number); if the square with the larger number of dots was selected (when the instruction was *Choose the square with few dots*); or if the square with the smaller number of dots was selected (when the instruction was *Choose the square with many dots*). No feedback was given on correct trials, but there was feedback on error trials in the form of the word "WRONG!!" which flashed on screen.

## Results

**Response times.** Response times (RTs) for trials with erroneous responses were discarded, leading to the loss of 354 trials from 5120, representing 6.9% of the trials. The correct response RTs were trimmed at 2.5 standard deviations for each subject, leading to the loss of a further 160 trials, or 3.4% of the remaining correct responses. Means for response times and error rates are given in Fig. (1). A linear mixed model of RT was built, using as independent variables *vagueness* and *numerical distance* and their interaction, with random slopes for *vagueness* and *numerical distance* over participants. *Vagueness* was sum coded: $vague = -.5$, $crisp = .5$; *numerical distance* was Helmert coded, i.e., the variable is ordered and each level is compared against the mean of all the previous levels to that point. All $p$ values were calculated using the R package *lmerTest* (Kuznetsova, Bruun Brockhoff, & Haubo Bojesen Christensen, 2014). Numbering below corresponds to the numbering of the hypotheses.

(1) RTs were faster for vague instructions than for crisp instructions ($\beta = .109$, $se = .022$, $t = 4.9$, $p < .001$).

(2) RTs grew faster as numerical distance increased: level one ($\beta = -.116$, $se = .012$, $t = -9.3$, $p < .05$), level 2 ($\beta = -.103$, $se = .009$, $t = -11.1$, $p < .001$) and level three ($\beta = -.082$, $se = .007$, $t = -11.0$, $p < .05$). Since discriminability of the arrays is easier for larger numerical distances, discriminability probably underlies this effect.

(3) Numerical distance and vagueness interacted significantly for larger numerical distances when modelling RT: essentially there were diminishing returns for vagueness as numerical distance increased.The interactions at the different levels of numerical distance were: level one: ($\beta = .001$, $se = .016$, $t = .03$, $p = .974$); level two ($\beta = -.039$, $se = .009$, $t = -4.460$, $p < .001$); level three ($\beta = -.043$, $se = .006$, $t = -7.0$, $p < .001$). In the crisp conditions RTs started out much slower than in the vague conditions, at the smallest numerical distance, but the two conditions converged to very fast times at the largest numerical distance.

**Error rates.** Error rate data were analysed using a generalized logit mixed model (Jaeger, 2008), with vagueness and numerical distance and their interaction as independent variables, and with random slopes for vagueness and numerical distance over participants.

(1) The effect of vagueness on error rates approached significance, with the vague instructions leading to fewer errors ($\beta = .307, se = .173, t = 1.8, p = .077$).

(2) Error rates decreased as numerical distance increased: level one ($\beta = -.585, se = .092, , z = -6.4, p < .001$), level 2 ($\beta = -.434, z = -5.3, p < .001$) and level three ($\beta = -.250, z = -4.1, p < .001$).

(3) Error rates were greater in the crisp conditions than the vague conditions when numerical distance was small, and this difference diminished with increasing numerical distance until it reversed at the biggest numerical distance.

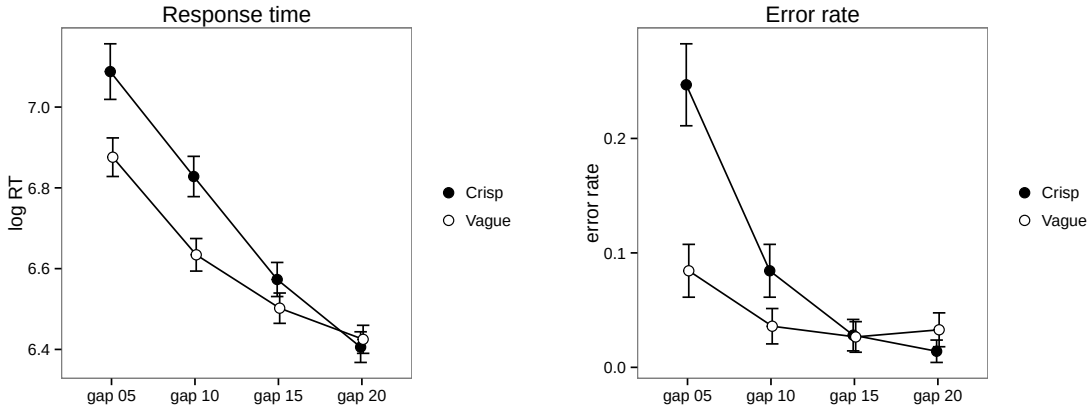


*Figure 1*. Experiment 1 results: response times and error rates

## Discussion

Responses were faster and more accurate for vague than crisp instructions. Response times and error rates improved as the arrays became more easily discriminable. The advantage for vague instructions tailed off as the arrays became more easily discriminable.

The cost reduction hypothesis explains the vagueness advantage by claiming that the vague referring expressions place less cognitive load on the comprehender than the crisp referring expressions. It explains the diminishing returns for vagueness in more-discriminable stimuli (i.e., the vagueness by numerical distance interaction) by claiming that load is low in both conditions for the easily-discriminable stimuli, and that therefore there is no extra benefit to be had from vagueness in the easily-discriminable stimuli.

# Experiment 2

## Introduction

The main result from experiment 1 was that responses were faster and more accurate for vague instructions than for crisp instructions.

Contrast, from Experiment 1, an expression from the vague condition: 'the square with few dots' with an expression from the crisp condition: 'the square with 5 dots'. One difference is that 'few' is vague (or at least has the potential for vagueness) and '5' is crisp. Another difference is that 'few' is a verbal quantifier while '5' is a numerical quantifier, in the sense that a number is mentioned explicitly. Since these two differences could not be separated in Experiment 1, the vagueness advantage finding is vulnerable to an alternative interpretation, that what we saw as a vagueness advantage was in contrast an advantage for the verbal form of the quantifier. In Experiment 2 we created verbal and numeric versions of each of the vague and crisp instructions so that we could compare vague and crisp conditions while taking account of verbal / numeric format.

In Experiment 1, the participants chose one of two squares: therefore the 'vague' quantifiers (e.g., 'few') uniquely identified one square. Recall our definition of vague – "a word is precise if it describes a well-defined set of objects. By contrast, a word is vague if it is not precise". In Experiment 1, it might be said that the quantifiers in the vague conditions did not realise their potential for vagueness. This is because there were no borderline cases of the referent that could make the referent set 'not well-defined', and perhaps because using definite articles in the instructions implied that only one option was correct. Using error feedback in experiment 1 could have exacerbated this. Experiment 2 used three arrays so that the vague quantifiers always had more than one possible referent, and used indefinite articles in the vague instructions to avoid the impression that only one response counted as correct, and was carried out without error feedback. An indication that the potential for vagueness was realised in Experiment 2 is that the borderline response was chosen fairly often: 16% of the time.

In Experiment 2, an item was a referring expression instruction followed by a set of dot arrays defined by triple of numbers, representing the number of dots in the left, middle, and right arrays. We used four different triples of numbers: (6,15,24); (16,25,34); (26,35,44); (36,45,54). Each set of arrays had the following properties: it comprised three arrays (instead of two as in Experiment 1); the array representing the central number was always presented in the middle of the three; there were two flanking arrays where one had fewer dots than the central array and the other had more.

Examples of crisp and vague versions of the numerical and verbal instructions follow: the examples assume the array (6,15,24) and reference to the smaller number of dots, such that 6 was classified as the expected response; 15 was classified as the borderline response; and 24 was classified as the incorrect response. In the *vague numerical* condition we used *Choose a square with about 10 dots*. None of the squares displayed contained 10 dots. 10 is slightly closer to 6 than to 15, justifying 6 as the best response and 15 as the borderline response. In the *vague verbal* condition we used *Choose a square with few dots*. In the *crisp numerical* condition we used *Choose the square with 6 dots* and one square always did contain the number mentioned. In the *crisp verbal* condition, we used *Choose the square with the fewest dots*.

**Hypotheses**

(1) A main effect RT advantage for vagueness.
(2) An RT advantage for vagueness both in *numeric* and in *verbal* instructions.
(3) No large main effect of instruction format.

(4) On the basis of Experiment 1, we would expect an effect of triple, with faster responses to more discriminable triples.

(5) Participants should make more borderline case choices for vague than crisp instructions.

## Method

Participants were aged between 18 and 45 with a median age of 28. We manipulated as independent variables vagueness and instruction format, yielding four conditions, *vague numeric*; *vague verbal*; *crisp numeric*; *crisp verbal*. We measured two dependent variables: response time; and the probability of a participant choosing the borderline case. On each trial, first the referring expression that constituted the instruction for that trial was displayed. Participants then pressed a key to indicate that they had read the instruction. After 1000 ms, the arrays were presented, while preserving the text of the referring expression. The response time dependent variable was measured from the presentation of the arrays, until the keypress indicating the participant's choice, which was also recorded. The trial would timeout after 60 seconds if there was no response. In this experiment, no feedback was given. This was because, in the vague conditions, we did not regard any response as 'correct' or 'incorrect', but instead as 'borderline response', or 'not borderline response', and we did not want to draw participants' attention to this distinction explicitly. We simply recorded whether the participant chose the best referent, the borderline case or the poorest referent, and how long it took the participant to respond.

## Results

**Response times.** Means for response times and proportion of borderline responses are given in Fig. (2). Response times from all trials were trimmed at 2.5 standard deviations for each subject, leading to the loss of 236 trials, 3.1% of the data. A linear mixed model was constructed for the (logged) response times. The *instruction format* and *vagueness* variables were sum-coded and *Item* was centred. The fixed effects in the model were *instruction format* and *vagueness* and their interaction, and *item*. The random effects in the model were *participant*, and slopes over *participant* for *instruction format*, *vagueness*, and their interaction

(1) The main effect of *vagueness* was to slow responses down, in contrast with Experiment 1, and offering evidence against hypothesis 1 (vague: 2668 ms; crisp: 2450 ms; a difference of 218 ms; $\beta = .06, se = .01, t = 4.6, p < .001$).

(2) Vagueness was disadvantageous in both the *numeric* and *verbal* conditions, offering evidence against hypothesis 2. The disadvantage for vagueness was greater in the numerical than in the verbal conditions, leading to an interaction effect between vagueness and instruction format ($\beta = -0.13, se = 0.02, t = -6.6, p = 0.000$).

(3) There was a significant effect of *instruction format* with numerical conditions attracting longer responses than the verbal conditions: consistent with Experiment 1 (numeric: 3284 ms; verbal 1866 ms; a difference of 1418 ms; $\beta = .37, se = .07, t = 5.1, p < .001$).

(4) There was a significant main effect of *triple* ($\beta = .06, se = .008, t = 7.0, p < .001$) indicating that response times differed in some way according to which triple was presented. However further analysis revealed that there was no consistent smooth trend across different

triples. This effect seems likely to be due to the very fast responses for the smallest triple, which had the largest ratio difference and so may have been particularly discriminable for participants.

**Borderline responses.** (5) Participant grand mean percentage of borderline selections was 16.6%. A generalized linear mixed model (Jaeger, 2008) was fit to the data for selection of the borderline response, with task, vagueness and item as fixed effects, and with random slopes for task and vagueness and item over participants. The distribution of responses over the nearest match square, the borderline square, and the furthest match square are given in Fig, 2. Participants were significantly more likely to choose the borderline option for vague instructions than for precise instructions (21.9% vs 11.3%, $\beta = .79, se = .25, z = 3.2, p < .01$). Participants were significantly more likely to choose the borderline square when the instruction used the numerical format rather than the verbal format (30.1% vs 3.0%, $\beta = 3.57, se = .26, z = 13.6, p < .001$).

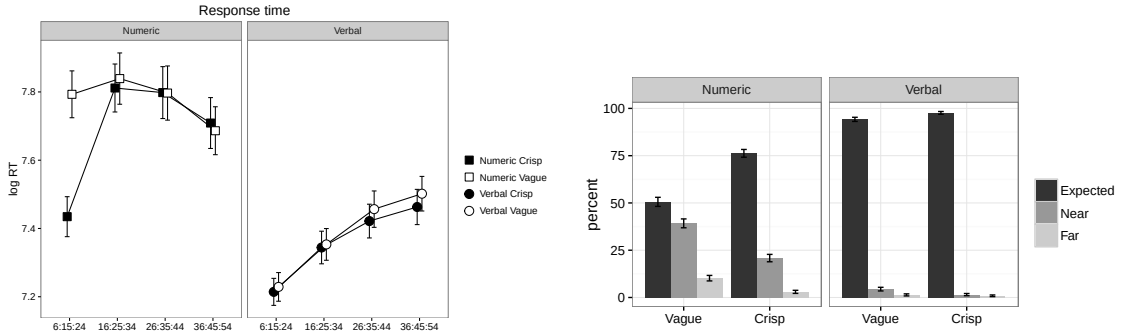


*Figure 2*. Experiment 2 results: response times and proportion of borderline cases

## Discussion

Experiment 2 tested to see whether when borderline cases are present, vague instructions would speed responses as they did in Experiment 1 when there were no borderline squares. We actually found a *disadvantage* of vague instructions: vague instructions slowed people down by 112 ms on average. We also found that the effect of instruction format was significant, with numerical format slowing responses by 689 ms on average, such that the disadvantage of numerical format overwhelmed the contribution of vagueness. The *verbal vague* condition was still responded to faster than the *numerical crisp* condition, so the pattern from Experiment 1 was reproduced, but in the light of the evidence from Experiment 2, in the presence of borderline cases, the advantage that was ascribed to vagueness before now looks more like an advantage of verbal instruction format.

Having effectively separated the **cost reduction** hypothesis from the **instruction format** hypothesis, it is important to observe that, in Experiment 2, instruction format went hand in hand with **selection algorithm**: the instructions that used a verbal instruction format allowed a comparison algorithm, whereas the instructions that used a numeric format allowed a matching algorithm. Therefore, our results so far permit the interpretation that what made the instructions in the verbal condition fast is not the fact that they were worded verbally, but that they allowed participants to use a comparison algorithm (which is known

to be faster than matching).

In the next two experiments we pitted the comparison algorithm and matching algorithm selection tasks against each other while controlling vagueness and instruction format. In Experiment 3 we restricted all the instructions to *numeric* quantifiers while factorially manipulating vagueness and selection task. In Experiment 4 we ensured that all instructions used *verbal* quantifiers, while also factorially manipulating vagueness and selection task. This allowed us to distinguish between the predictions of the selection task account and the instruction format account.

## Experiment 3

### Introduction

The main aim of experiment 3 was to see whether vagueness would exert beneficial effects when all conditions used numerals in the instructions, and when there were vague and crisp versions of the instructions for both comparison and matching strategies. The main changes from experiment 2 were that the selection task was explicitly controlled, and that all conditions were constrained to mention a number. We used the same arrays as in experiment 2. Table 1 shows the instructions for each condition.

Table 1
*Experiment 3: Instructions, assuming* $6, 15, 24$ *dots as the Item, and showing* fewer *instead of* more

| vagueness | instruction format | selection task | instruction |
|---|---|---|---|
| crisp | numeric | matching | Choose a square with 6 dots |
| crisp | numeric | comparison | Choose a square with fewer than 20 dots |
| vague | numeric | matching | Choose a square with about 10 dots |
| vague | numeric | comparison | Choose a square with far fewer than 20 dots |

### Hypotheses

(1) Vague instructions are easier for the reader than crisp alternatives (main effect of vagueness)

(2) Comparison is easier for the reader than matching (main effect of selection task)

(3) Effects of vagueness are different depending on whether selection is matching or comparison (interaction effect selection x vagueness).

### Method

38 participants were recruited. The design was a 2 x 2 factorial manipulation of vagueness and selection task (see Table 1). On each trial a referring expression instruction was presented: participants pressed a key to dismiss the instruction, when the dot arrays were presented until the participant responded, and the response time and choice were recorded.

## Results

A linear mixed effects regression model was built for log response times. The structure of the model was as follows: fixed effects were vagueness, selection task (both sum coded) and centred item and their interactions: random effects were vagueness, selection task, and item (but not their interactions - the model failed to converge when these interactions were included). The means are plotted in Figure 3.

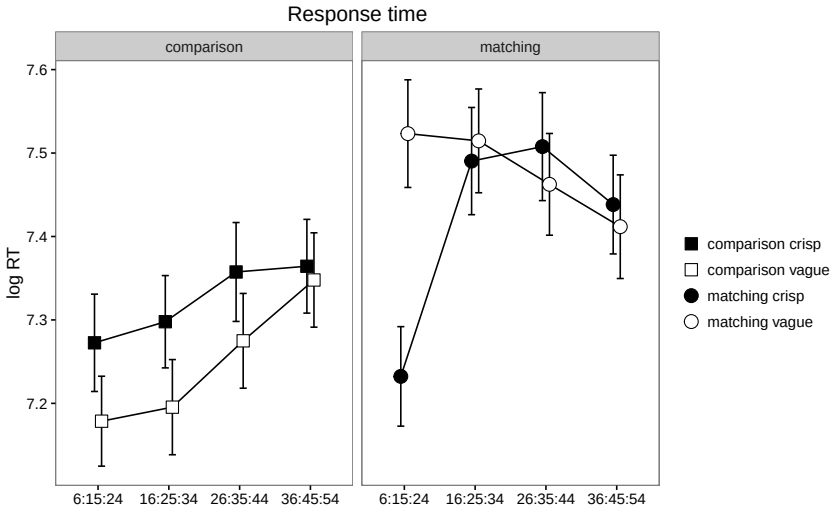


*Figure 3*. Results for Experiment 3

(1) There was no significant main effect of vagueness ($\beta = .003, se = .014, t = .202, p = .841$). The results showed that vagueness was beneficial for comparison but detrimental for matching. (2) There was a main effect of selection task, with the comparison task speeding responses compared to the matching task ($\beta = -.165, se = .027, t = -6.218, p < .001$). (3) Vagueness exerted effects in different directions for the comparison task and for the matching task. Separate analyses were conducted at each level of the selection task to see whether within each task type there were significant effects of vagueness. There were: in the comparison task vagueness significantly speeded response times compared with crisp controls ($\beta = -.07, se = .02, t = 3.52, p < .01$). In the matching task vagueness significantly slowed response times compared with crisp controls ($\beta = -.07, se = .02, t = -2.89, p < .05$).

None of the accounts set out in the Hypotheses section emerge well from the results. The instruction format account wrongly predicts no differences between the conditions. The selection task correctly predicted the main effect of selection task, but has no coverage of the interaction with vagueness. The cost reduction account is wrong to predict main effect advantages for vagueness, and wrong to predict that vagueness should be beneficial at each level of the selection task: however vagueness was advantageous in the comparison task.

## Experiment 4

### Introduction

This experiment investigated response times for instructions that did not use a number. We manipulated vagueness and the selection task (comparison and matching). In order to implement the experiment without mentioning numbers in the instructions, we changed the sequence of each trial to include a 'target' (i.e., a dot array of a particular cardinality) before the instruction, so that we could then refer back to the target's cardinality in the instruction using expressions like *the same number of dots as the target*; *fewer dots than the target*. This presentation of a target before the main body of the trial shares some features with Izard and Dehaene (2008, Experiment 2), although in that experiment participants were told the cardinality of the target (called an *inducer* in that paper) whereas in our experiment we did not tell participants the cardinality of the prime array. An item was thus a combination of a target dot array, an instruction that did not explicitly contain a number, and a set of dot arrays taking their cardinalities from the same triples used in experiment 2. Table 2 spells out how the instructions were constrained not to mention a numeral.

Table 2
*Experiment 4: Instructions*

| vagueness | instruction format | selection task | instruction |
|---|---|---|---|
| crisp | verbal | matching | Choose a square with the same number of dots as the target |
| crisp | verbal | comparison | Choose a square with fewer dots than the target |
| vague | verbal | matching | Choose a square with about the same number of dots as the target |
| vague | verbal | comparison | Choose a square with far fewer dots than the target |

### Hypotheses

(1) Vague instructions are easier for the reader than crisp alternatives (main effect of vagueness)

(2) Comparison is easier for the reader than matching (main effect of selection)

(3) Effects of vagueness are different depending on whether selection is matching or comparison (interaction effect selection x vagueness).

### Results

40 volunteers participated. The results showed that vagueness was beneficial for comparison but detrimental for matching (the same as Experiment 3) even when no numbers were allowed in the instructions. Figure 4 shows the means by condition.

(1) There was no main effect of vagueness ($\beta = .01, se = .01, t = 1.51, p = .14$).

(2) There was a main effect of selection, with comparison task instructions leading to faster responses than the matching task instructions ($\beta = -.18, se = .02, t = -10.38, p < .01$). This effect was in the same direction as Experiment 3.

(3) Vagueness did exert different effects depending on the selection task ($\beta = .12, se = .03, t = 4.32, p < .05$). Separate analyses were conducted for the comparison task and for the matching task. In the comparison task, vagueness resulted in faster response times ($\beta = -0.08, se = .02, t = 4.30, p < .05$). In the matching task vagueness slowed response times ($\beta = .05, se = .01, t = 3.72, p < .05$). These results are in the same direction as Experiment 3.

The cost reduction account is wrong to predict main effect advantages for vagueness, and wrong to predict that vagueness should be beneficial at each level of the selection task: however vagueness was advantageous in the comparison task. The instruction format account wrongly predicts no differences between the conditions. The selection task account correctly predicted the main effect of selection task, but has no coverage of the interaction of selection task with vagueness.

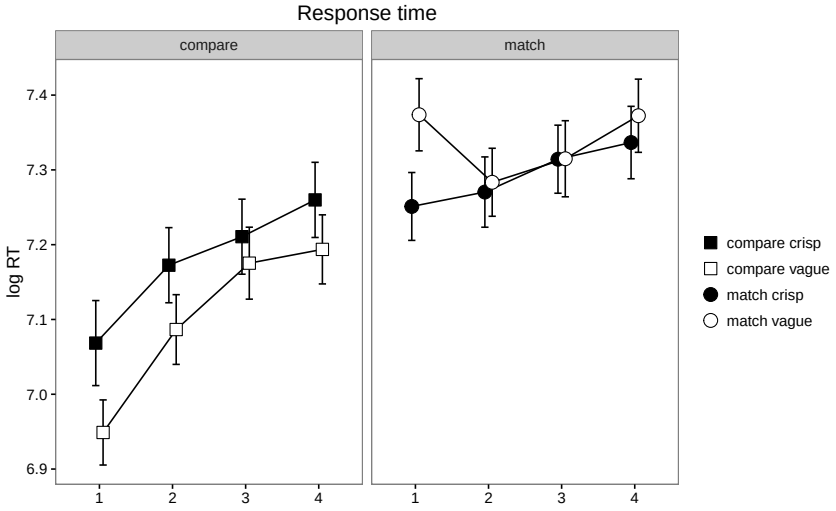


*Figure 4*. Results for Experiment 4

## Discussion of experiments 3 and 4

The main aim of these two experiments was to test whether vagueness confers any cognitive benefits over and above those due to differences in the selection task according to whether the instruction mandates a *comparison* selection task or a *matching* selection task, when number-use is held constant. The main effect of selection task showed that the assumption that the *comparison* task is easier than the *matching* task is well-founded. In both experiments people were reliably quicker to respond in the *comparison* task.

Vagueness, which was the phenomenon on which our investigation focussed, did not exert a main effect in response time. However when the comparison and selection tasks were analysed separately, there was small reliable advantage for vagueness in the *comparison* tasks, but a small reliable disadvantage for vagueness in the *matching* tasks.

## General Discussion

Experiment 1 showed us that responses were faster and more accurate when the instructions were vague than when they were crisp, but the experiment could not distinguish effects of vagueness from those of number-avoidance or selection task: the vague conditions were also in verbal rather than numerical format; and mandated a comparison strategy rather than a matching strategy. Experiment 2 showed us that number avoidance in the verbal format instructions is an important factor driving the faster response times in the task, and that vagueness does not have any additional explanatory power in either the verbal format instructions or the numerical format instructions when we generated verbal and numerical versions of both crisp and vague instructions. However the experiment could not distinguish benefits of number-avoidance in the verbal instructions from benefits of the comparison selection task: the verbal instructions also mandated a comparison strategy rather than a matching strategy. In experiments 3 and 4 we manipulated vagueness and the selection task independently of numerical format. We found that there are effects of the selection task mandated by the instruction, with the comparison task instructions attracting faster response times than the matching instructions, and that vagueness exerts benefits when the selection task is *comparison*, but not when the task is *matching*.

The benefits of vagueness in the *comparison* task in experiments 3 and 4 could be explained as differences in the number of valid targets for the expression, as follows. Taking as an example the stimulus with (6,15,24) dots, it could be argued that the vague comparison instruction (e.g., *a square with far fewer than 20 dots*) has one valid target, the square with 6 dots, while the crisp comparison instruction (e.g., *a square with fewer than 20 dots*) has two valid targets, the squares with 6 and 15 dots. In both experiments 3 and 4 we found that people were quicker to identify a square when the instruction only had one valid target. This leads us to speculate that the benefit for vagueness here could be due to the vague expression foregrounding a particular valid target while the crisp expression carries with it the additional task of distinguishing between two alternative valid targets, something we propose to call a "range-reduction" benefit.

What is one entitled to conclude? Given that we were able to identify a class of situations – namely: situations in which a comparison strategy suffices to identify the intended referent – in which vague expressions led to faster response times than crisp ones, would it be valid to conclude that we have finally discovered an advantage for vagueness that cannot be ascribed to some other factor? We believe the answer to this question is negative. To see why, consider figures 3 and 4. Both figures depict four conditions, depending on whether the expression was crisp or vague, and depending on whether the referent could be identified using a comparison strategy or not. Two of the resulting four conditions result in an expression that can denote either of two referents; the other two conditions result in an expression that can only denote one referent, with the other possible referent being a marginal candidate at best:

To see why vagueness thus has opposite effects, depending on whether it is used in matching or comparison situations, compare an instruction like 'Choose a square with 6 dots' with its vague counterpart 'Choose a square with about 10 dots': by adding the word 'about', we broaden the range of squares that the expression might be referring to. On the other hand, compare 'Choose a square with fewer than 20 dots' with its vague counterpart

Table 3
*Vagueness as range reduction*

| vagueness | selection task | candidates |
|---|---|---|
| crisp | matching | 1 candidate |
| crisp | comparison | 2 candidates |
| vague | matching | 2 candidates |
| vague | comparison | 1 candidate |

'Choose a square with far fewer than 20 dots': by adding the word 'far', we did not broaden the range of squares denotable by the expression: we narrow it down, because only some of the squares that have fewer dots may have *far* fewer dots.

The observation that conditions with 1 candidate lead to shorter response times than conditions with 2 candidates is consistent with the range reduction hypothesis, but not with the idea that vagueness has a beneficial effect. It appears, in other words, that range reduction causes shorter response times, suggesting that shorter response times will only result from a vague expression if this expression leads to range reduction. Once again, it seems, it is not vagueness itself that has advantages but a phenomenon (namely range reduction) that is an automatic concomitant of vagueness in some types of situations.

The findings from our experiments show that when vague expressions are compared with crisp alternatives in our forced choice task, vague expressions appear to yield benefits in some situations, but that the observed benefits may be due to factors other than vagueness itself that the vague forms bring along with them: factors like avoiding numbers; permitting comparison tasks; and range reduction. The picture that is starting to emerge, in other words, is subtle: on the one hand, in the situations that we have been studying – where cooperative speakers refer to an object (e.g., a square) by means of some quantity associated with the object (e.g., the number of dots contained in the square) – vagueness is not intrinsically beneficial. On the other hand, vague expressions often have other features that *are* beneficial, and these are what give us the incorrect impression that vagueness itself is beneficial. Vagueness may thus have acquired a reputation that it does not deserve.

A comparison may clarify the logic of the situation. In recent years a number of studies, focussing typically on red wine, have suggested that alcohol, consumed in low doses, may have health benefits. An alternative explanation, however, asserts that it is not the alcohol in the wine that was beneficial, but antioxidants from grapes. If this alternative explanation is correct then alcohol may not be as beneficial as some would like to think.

**Implications for practical NLG systems.** Our findings suggest a re-think of the questions on which much research on the utility of vagueness rests. The question of how a particular piece of quantitative information is best conveyed through language is certainly an important one for practical NLG. Years of research on the logic of vagueness in natural language have primed the research community to expect that the utility of vagueness is an important part of the answer, but our findings call this into question.

Consider, once again, a number of ways in which a given temperature can be reported. Given a temperature measurement (or prediction) as input, an NLG system might justifiably

say that it was

(a) *27.2 degrees*, or
(b) *approximately 27 degrees*, or
(c) *above 25 degrees*, or
(d) *warm*,

among other candidate expressions. If standard accounts are to be believed, then options (a) and (c) convey crisp information, whereas (b) and (d) are vague (i.e., they permit borderline cases). But, to the best of our knowledge, there is no experimental evidence that the fact in itself that an expression is vague confers a benefit (or a de-benefit) on it for hearers. Rather than asking whether a candidate expression is vague, other questions might shed more light on the NLG system's choice, similar (though not necessarily identical) to the ones identified in our empirical studies. For a start, both (c) and (d) designate an interval that is half-bounded, whereas both (a) and (b) express closed intervals, and this might affect their comprehension by a hearer (cf., our finding about comparison versus matching strategies). Furthermore, the expressions (c) and (d) – one of which is vague while the other is crisp – appear to convey very similar amounts of information, saying that the temperature is higher than some (fairly high) standard. Perhaps most importantly, "warm" is shorter than the other three expressions and avoids the use of numbers, and our experiments suggest that this may make "warm" more rapidly understood than its competitors; earlier experiments point in the same direction, given the evaluative nature of the word "warm" (see section 1 and Peters et al. (2009)).

One way to see why vagueness may not matter as much for NLG – and for human communication more generally – as is often thought is the following thought experiment. Many speakers understand the word "warm" as vague. Now suppose a group of users agreed to give it a precise definition; according to this precisification, "warm" means $> 25$ degrees (as in (c)). It seems unlikely that this change, from a vague meaning to a crisp one, would change the utility of the word.

## References

Allen, R. (2000). *The New Penguin English Dictionary*. Penguin Books.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436.

Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. *Econometrica*, *50*(6), 1431–1451.

Dehaene, S. (1996). The organization of brain activations in number comparison: Event-related potentials and the additive-factors method. *Journal of Cognitive Neuroscience*, *8*(1), 47-68. (cited By (since 1996) 152)

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General*, *122*(3), 371.

De Jaegher, K. (2003). A Game–Theoretic Rationale for Vagueness. *Linguistics and Philosophy*, *26*, 637–659.

Eaton, J. W. (2002). GNU *Octave Manual*. Network Theory Limited.

Egre, P., & Klinedinst, N. (2011). Introduction: Vagueness and language use. In P. Egre & N. Klinedinst (Eds.), *Vagueness and Language Use*. Palgrave.

Gevers, W., Lammertyn, J., Notebaert, W., Verguts, T., & Fias, W. (2006). Automatic response activation of implicit spatial information: Evidence from the SNARC effect. *Acta Psychologica*, *122*(3), 221–233.

Goldberg, E., Driedger, N., & Kittredge, R. (1994). Using natural-language processing to produce weather forecasts. *IEEE Expert*, *9*(2), 45–53.

Hobbs, J. R. (1985). Granularity. In *In proceedings of the ninth international joint conference on artificial intelligence* (pp. 432–435). Morgan Kaufmann.

Hripcsak, G., Elhadad, N., Chen, Y., Zhou, L., & Morrison, F. P. (2009). Using Empiric Semantic Correlation to Interpret Temporal Assertions in Clinical Texts. *Journal of the American Medical Informatics Association*, *16*(2), 220-227. doi: 10.1197/jamia.M3007

Hunter, J., Freer, Y., Gatt, A., Logie, R., McIntosh, N., Van Der Meulen, M., ... Sykes, C. (2008). Summarising complex ICU data in natural language. In *AMIA Annual Symposium Proceedings* (Vol. 2008, p. 323). American Medical Informatics Association.

Izard, V., & Dehaene, S. (2008). Calibrating the mental number line. *Cognition*, *106*(3), 1221 - 1247.

Jaeger, T. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

Keefe, R., & Smith, P. (Eds.). (1996). *Vagueness: a Reader. A Bradford Book*. The MIT Press.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception*, *36*.

Kuznetsova, A., Bruun Brockhoff, P., & Haubo Bojesen Christensen, R. (2014). lmertest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=lmerTest` (R package version 2.0-6)

Lipman, B. L. (2000). "Comments section". In A. Rubinstein (Ed.), *Economics and language: Five essays*. Cambridge Univ Press.

Lipman, B. L. (2009). *Why is Language Vague?* (retrieved 12 April 2011 from `http://people.bu.edu/ blipman/Papers/vague5.pdf`)

Mishra, H., Mishra, A., & Shiv, B. (2011). In Praise of Vagueness: Malleability of Vague Information as a Performance-Booster. *Psychological Science*.

Peters, E., Dieckmann, N., Västfjäll, D., Mertz, C., Slovic, P., & Hibbard, J. (2009). Bringing meaning to numbers: The impact of evaluative categories on decisions. *Journal of Experimental Psychology: Applied*, *15*(3), 213.

Portet, F., Reiter, E., Gatt, A., Hunter, J., Sripada, S., Freer, Y., & Sykes, C. (2009). Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, *173*(7-8), 789–816.

Reiter, E., & Dale, R. (2000). *Building natural language generation systems*. Cambridge University Press.

Turner, R., Sripada, S., Reiter, E., & Davy, I. (2006). Generating spatio-temporal descriptions in pollen forecasts. In *EACL '06: Proceedings* (pp. 163–166). Stroudsburg, PA: Association for Computational Linguistics.

van Deemter, K. (2009). Utility and Language Generation: The Case of Vagueness. *Journal of Philosophical Logic*, *38*(6), 607–632.

van Deemter, K. (2010). Vagueness Facilitates Search. In *Logic, Language and Meaning: 17th Amsterdam Colloquium, Amsterdam, the Netherlands, December 16-18, 2009, Revised Selected Papers* (p. 173). New York, NY: Springer-Verlag New York Inc.

van Rooij, R. (2003). Being polite is a handicap: Towards a game theoretic analysis of polite linguistic behavior. In M. Tenneholz (Ed.), *TARK 9: Theoretical Aspects of Rationality and Knowledge*. Bloomington: Bloomington.

van Oeffelen, M., & Vos, P. (1982). A probabilistic model for the discrimination of visual number. *Perception and Psychophysics*, *32*(2), 163–170.