Is Vagueness Beneficial for Hearers?

Evidence from Experiments

Matt Green

Computing Science

University of Aberdeen

Kees van Deemter

Computing Science

University of Aberdeen

Author Note

latex compiled Tuesday 27th September, 2016 13:32

Abstract

Much of everyday language use 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 to let the system generate vague expressions. The present paper is an attempt to find out what benefits vagueness might have for readers.This article 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 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. The paper concludes with a brief discussion of the implications for practical NLG.

Is Vagueness Beneficial for Hearers?
Evidence from Experiments

1 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 & 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

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

"vague predicates have borderline cases, have fuzzy boundaries, and are susceptible to sorites paradoxes" (Keefe & Smith, 1997, 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, 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 weight to a computer each week. Participants were told that two algorithms would be used to compute their HHI, and that the two might give different values initially, in which case 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, 1997; 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*.

1.1 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.

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 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 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. All p values reported for linear models were calculated using the R package *lmerTest* (Kuznetsova, Bruun Brockhoff, & Haubo Bojesen Christensen, 2016).

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. 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.

There is abundant evidence (e.g., Trick & Pylyshyn, 1994) that very small (i.e., *subitizable*) quantities are recognised and processed by a distinct psychological mechanism that differs from that used to process larger quantities. We performed a pilot experiment (Green & van Deemter, 2011) in which we were able to confirm this finding in the

experimental settings on which we are focussing in this paper. As one might expect, we found that, when participants were confronted with a stimulus consisting of two squares containing different numbers of dots[2], instructions of the form *Choose the square with* n *dots* led to consistently faster response times than instructions of the form *Choose the square with many/few dots* when $2 \leq n \leq 5$; the converse was true for $n > 5$. We henceforth focussed our studies on non-subitizable numbers, because it is there that vagueness is expected to have benefits.

2 Experiment One

2.1 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 dot arrays was presented, each in the form of a square containing some number of dots. The participant was required to identify the dot 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

[2]such a stimulus is referred to hereafter as a stimulus consisting of a set of some number of dot arrays, where the number of dots in each dot array varies and is referred to as its cardinality, and the physical arrangement of dots in each dot array is irregular

cardinality, the *crisp* instruction was *Choose the square with 5 dots* and its *vague* counterpart was *Choose the square with few dots*.

2.2 Hypothesis

(H1) A main effect advantage for vagueness: vague instructions impose a lower cognitive load for the comprehender than crisp alternatives.

2.3 Method

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. Hypothesis H1 was tested looking both at response times and at error rates.

2.4 Results

2.4.1 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

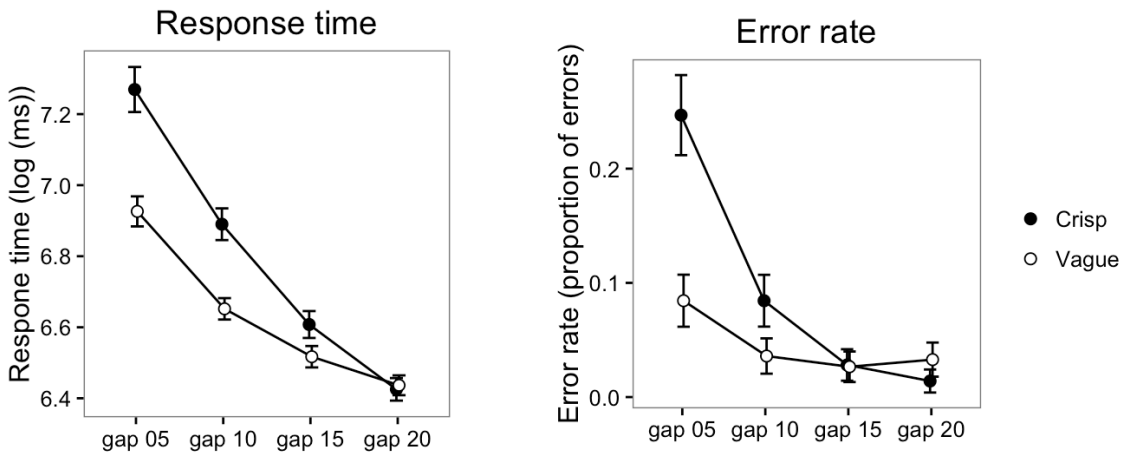

Figure 1. Experiment 1 results: response times and error rates.

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, with vagueness and numerical distance and their interaction as independent variables, and with random slopes for vagueness and numerical distance over participants.

($H1_{rt}$) RTs were faster for vague instructions than for crisp instructions ($\beta = -0.11$, $se = .02$, $t = 4.7$, $p < .001$).

2.4.2 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.

($H1_{er}$) There was a statistically non-significant effect of vagueness on error rates: the vague instructions tended to have a lower error rate than the than the crisp instructions ($\beta = -.22$, $se = .18$, $t = -1.2$, $p = .230$).

2.5 Discussion

As expected, responses were faster and more accurate for vague than crisp instructions. Post hoc, we formulated some further hypotheses:

(H2) 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.

(H3) 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.

Looking at response times, these two post-hoc hypotheses were supported:

($H2_{rt}$) RTs increased with numerical distance ($\beta = -0.18$, $se = 0.01$, $t = -11.6$, $p < 0.0005$). Since discriminability of the arrays is easier for larger numerical distances, discriminability probably underlies this effect.

($H3_{rt}$) Numerical distance and vagueness interacted significantly: essentially there were diminishing returns for vagueness as numerical distance increased: ($\beta = 0.07$, $se = 0.01$, $t = 5.9$, $p < 0.0005$).

Looking at error rates, the two post-hoc hypotheses were supported:

($H2_{er}$) Error rates decreased as numerical distance increased: ($\beta = -0.78$, $se = 0.08$, $z = -9.6$, $p < 0.0005$).

($H3_{er}$) Numerical distance and vagueness interacted significantly: essentially there were diminishing returns for vagueness as numerical distance increased until the biggest numerical distance when vagueness was disadvantageous: ($\beta = 0.80$, $se = 0.15$, $z = 5.27$, $p < 0.0005$).

So far, these results are in line with the idea of cost reduction. Cost reduction 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.

3 Experiment 2

3.1 Introduction

However, the picture painted by these findings might be misleading. First of all, there is a possibly confounding factor. 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' has the potential for vagueness, whereas '5' is crisp. But another difference is that 'few' is verbal while '5' is numerical, 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 therefore 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.

Another potential problem with Experiment 1 is the following. 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, 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.

To find out what happens when words are used in a context where their potential for vagueness comes to the fore, 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 three dot arrays defined by a 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 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. For *crisp verbal*, we used *Choose the square with the fewest dots*.

Table 1

Experiment 1: instructions arranged by condition for the dot triple $(6, 15, 24)$ *and instructions indicating the smaller numbers*

	crisp	vague
numerical	Choose the square with 6 dots	Choose a square with about 10 dots
verbal	Choose the square with the fewest dots	Choose a square with few dots

3.2 Hypotheses

(H1) A main effect RT advantage for vagueness.

(H2) An RT advantage for vagueness both in *numeric* and in *verbal* instructions.

(H3) No large main effect of instruction format (of numeric versus verbal) since we hypothesised that vagueness rather than instruction format drove the effect in experiment 1.

(H4) On the basis of Experiment 1, we would expect faster responses for stimuli with more discriminable arrays.(i.e., an effect of item).

(H5) Participants should make more borderline case choices for vague than crisp instructions.

3.3 Method

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.

3.4 Results

3.4.1 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, with sum-coded vagueness, instruction format, (and their interaction), and item as fixed effects, and the same effects as slopes over participant for random effects.

(H1) 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; $\beta = .06$, $se = .01$, $t = 4.6$, $p < 0.0001$)

(H2) In focussed comparisons, vagueness was significantly disadvantageous in the *numeric* conditions ($\beta = 0.09$, $se = 0.02$, $t = 4.5$, $p < 0.0001$), and non-significantly disadvantageous in the *verbal* conditions ($\beta = 0.02$, $se = 0.02$, $t = 1.5$, $p = 0.1550$), offering partial evidence against hypothesis 2. The disadvantage for vagueness was greater in the numerical than in the verbal conditions, leading to a significant overall interaction effect between vagueness and instruction format ($\beta = -0.13$, $se = 0.02$, $t = -6.6$, $p = 0.0169$).

(H3) There was a significant effect of *instruction format* with numerical conditions attracting longer responses than the verbal conditions: consistent with Experiment 1, but suggesting that instruction format rather than vagueness drove the effect we observed in experiment 1 (numeric: 3284 ms; verbal 1866 ms; a difference of 1418 ms; ($\beta = -0.36$, $se = 0.07$, $t = -5.1$, $p < 0.0001$).

(H4) There was a significant main effect of item (i.e., of which triple of numbers of dots were used in the stimulus): ($\beta = 0.12$, $se = 0.02$, $t = 7.1$, $p < 0.0001$). This effect seems likely to be due to the very fast responses for stimuli using arrangements of the smallest numbers of dots (6,15,24), which also had the largest difference in ratio of smallest number to largest number in the stimulus, suggesting that stimuli using these numbers of dots may have been particularly discriminable for participants.

3.4.2 Borderline responses. A generalized linear mixed model (Jaeger, 2008) was fit to the data for selection of the borderline response, with sum-coded vagueness, instruction format, (and their interaction), and item as fixed effects, and the same effects

as slopes over participant for random effects. The distribution of responses over the nearest match square, the borderline square, and the furthest match square are given in Fig, 2. Participants chose the borderline square on 16.6% of trials overall.

(H5) Participants were significantly more likely to choose the borderline option for vague instructions than for crisp instructions (21.9% vs 11.3%: $\beta = 0.62$, $se = 0.22$, $z = 2.8$, $p = 0.0059$). Participants were also 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%: $\beta = -3.35$, $se = 0.23$, $z = -14.6$, $p < 0.0001$).

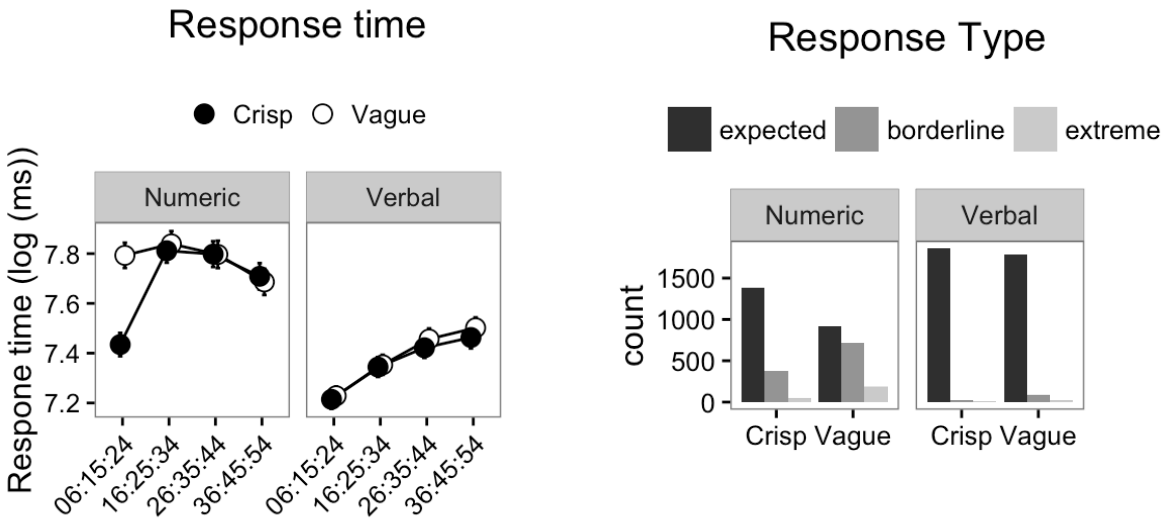

Figure 2. Experiment 2 results: mean response times by condition and item, and counts of borderline case responses by condition.

3.5 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.

However, once again there is a possibly confounding factor. Observe that, in Experiment 2, instruction format (i.e., the choice between numeric and verbal) went hand in hand with might be called the (human) **selection algorithm**: To see this, consider the task of selecting the dot array that contains “few dots": to do this, it suffices to *compare* the three arrays and select the one that contains the fewest elements. 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). **–Insert citation**

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.

4 Experiment 3

4.1 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 2 shows the instructions for each condition.

KvD I think we should motivate our choice of materials (especially "far ..er")

Table 2

Experiment 3: Instructions arranged by condition for stimuli with $(6, 15, 24)$ *dots and instructions indicating the smaller numbers*

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

4.2 Hypotheses

(H1) Vague instructions are easier for the reader than crisp alternatives (main effect of vagueness)

(H2) Comparison is easier for the reader than matching (main effect of selection task)

(H3) Effects of vagueness are different depending on whether selection is matching or comparison (interaction effect selection x vagueness).

4.3 Method

38 participants were recruited. The design was a 2 x 2 factorial manipulation of vagueness and selection task (see Table 2). 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.

4.4 Results

Response times were trimmed at 2.5 SD separately for each subject, leading to the loss of 204 trials (2.8% of the trials). Condition means for the remaining (logged) RTs are plotted in Figure 3. A linear mixed model was constructed for the logged response times, with sum-coded vagueness, instruction format, (and their interaction), and item as fixed effects, and the same effects as slopes over participant for random effects.

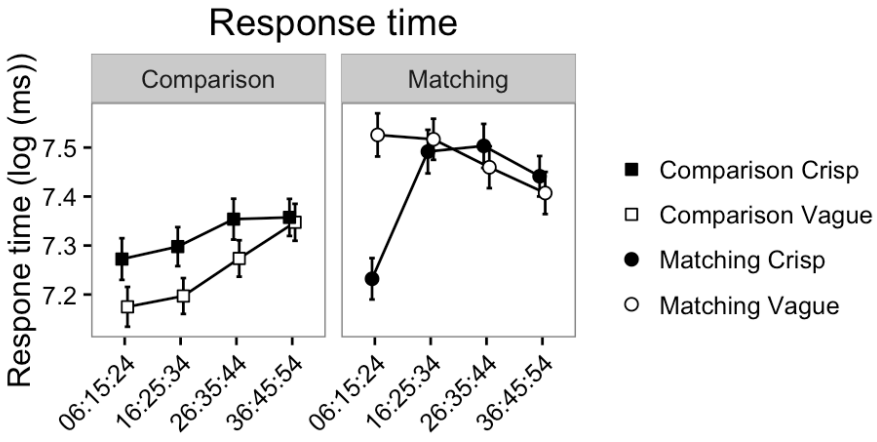

Figure 3. Mean response times by condition for Experiment 3 where all instructions were numeric

(H1) Vague instructions were non-significantly easier for the reader than crisp instructions when vagueness was considered as a main effect ($\beta = -0.01$, $se = 0.01$, $t = -0.4$, $p = 0.7158$).

(H2) There was a statistically significant main effect of selection task, with the comparison task speeding responses compared to the matching task ($\beta = 0.16$, $se = 0.03$, $t = 6.2$, $p < 0.0001$).

(H3) Vagueness exerted effects in different directions for the comparison task and for the matching task: there was a significant interaction between vagueness and selection task ($\beta = 0.13$, $se = 0.03$, $t = 4.2$, $p < 0.0001$). Separate analyses were conducted testing for effects of vagueness at each level of the selection task. Within the comparison task

vagueness significantly speeded response times compared with crisp controls ($\beta = -0.07$, $se = 0.02$, $t = -3.5$, $p < 0.0012$). Within the matching task vagueness significantly *slowed* response times compared with crisp controls ($\beta = 0.06$, $se = 0.02$, $t = 2.9$, $p < 0.0061$).

The cost reduction account was wrong to predict significant main effect advantages for vagueness (although there was a non-significant trend in the *direction* predicted by the cost reduction account), and wrong to predict that vagueness should be beneficial at each level of the selection task: however vagueness was significantly advantageous in the comparison task.

5 Experiment 4

5.1 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 contain a number, and a set of dot arrays taking their cardinalities from the same triples used in Experiment 2. Table 3 spells out how the instructions were constrained not to mention a numeral and gives examples of targets.

Table 3

Instructions and targets by condition for experiment 4

Item	Selection	Vagueness	Target	Choose a square with...
06:15:24	Comparison	Crisp	20	...fewer dots than the target
		Vague	20	...far fewer dots than the target
	Matching	Crisp	6	...the same number of dots as the target
		Vague	10	...about the same number of dots as the target

5.2 Hypotheses

(H1) Vague instructions are easier for the reader than crisp alternatives (main effect of vagueness)

(H2) Comparison is easier for the reader than matching (main effect of selection)

(H3) Effects of vagueness are different depending on whether selection is matching or comparison (interaction effect selection x vagueness, and focussed comparisons at each level of selection).

5.3 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.

(H1) There was no significant main effect of vagueness ($\beta = -0.02$, $se = 0.01$, $t = -1.5$, $p = 0.1296$).

(H2) There was a main effect of selection, with comparison task instructions leading to faster responses than the matching task instructions ($\beta = 0.18$, $se = 0.02$, $t = 10.4$, $p < 0.0001$). This effect was in the same direction as Experiment 3.

(H3) Vagueness did exert different effects depending on the selection task (main

interaction effect of vagueness by selection $\beta = 0.12$, $se = 0.02$, $t = 5.1$, $p < 0.0001$). Separate analyses of the effect of vagueness were conducted for the comparison task and for the matching task using Bonferroni-adjusted significance thresholds. In the comparison task, vagueness resulted in faster response times ($\beta = -0.08$, $se = 0.02$, $t = -4.3$, $p < 0.0001$). In the matching task vagueness slowed response times ($\beta = 0.05$, $se = 0.01$, $t = 3.7$, $p = 0.0004$). These results are in the same direction as Experiment 3.

The cost reduction account was 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.

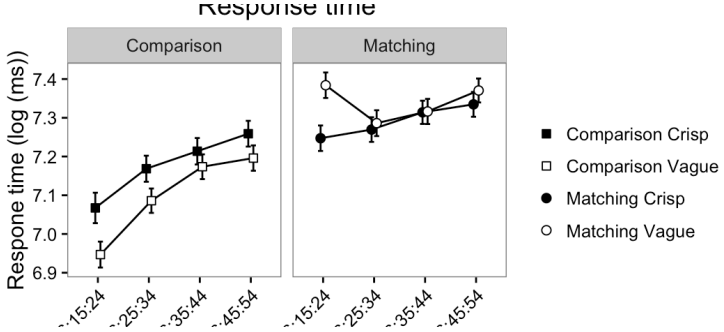

Figure 4. Mean response times by condition for Experiment 4

6 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 significant main effect in response time. However when the comparison and selection tasks were analysed separately, there was small significant advantage for vagueness in the

comparison tasks, but a small significant disadvantage for vagueness in the *matching* tasks.

7 General Discussion

Experiment 1 showed 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 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 benefit in either the verbal format instructions or the numerical format instructions.

However, Experiment 2 could not distinguish benefits of number avoidance from benefits of the comparison selection task. In Experiments 3 and 4 we manipulated vagueness and the selection task separately at each level of numerical format. Across the two experiments, we found that the comparison-task instructions attracted faster response times than the matching-task instructions. Within the two experiments we found 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.

Table 4

Vagueness as range reduction: a summary of Experiments 3 and 4

selection task	vagueness	candidates	effect of vageueness
comparison	crisp	2	vagueness advantage
	vague	1	
matching	crisp	1	vagueness disadvantage
	vague	2	

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.

KvD Something missing here?

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

'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 is beneficial. It appears, in other words, that shorter response times will only result from a vague expression if this expression leads to range reduction. Once again, 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.

Our findings suggest that the observed benefits of vague expressions in certain situations may be due to factors other than vagueness itself: factors like avoiding numbers; permitting comparison tasks; and range reduction. The picture that is starting to emerge 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 – vagueness is not intrinsically beneficial. On the other hand, vague expressions frequently possess 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. To answer to Lipman's question, of why vagueness permeates human language (see our Introduction), may lie in a different direction after all, possibly relating to a combination of pure necessity and benefits for the speaker rather than the hearer (see van Deemter, 2010, chapter 11, for discussion of some possible directions).

A comparison may clarify the logic of the situation. In recent years a number of studies, focussing 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 healthy after all.

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 an important one for practical NLG. Years of research on the logic of vagueness in natural language – which has given rise to such logical techniques as Partial Logic (e.g, (Fine, 1975), Probabilistic Logic (Edgington, 1997), and Fuzzy Logic (Zadeh, 1965) – have primed the research community to expect that the utility of vagueness is an important part of the answer, but our findings call this expectation into question. Although our own studies in this article have focussed on vagueness in descriptive Noun Phrases, it seems plausible that vagueness plays a similar role in other linguistic constructs.

For example, consider air temperature, once again. Given a temperature measurement (or prediction) as input, an NLG system that addresses a non-expert audience might say that it was (or will be)

(a) *27.2 degrees Celsius*, or

(b) *approximately 27 degrees*, or

(c) *above 25 degrees*, or

(d) *warm*,

among other candidate expressions. If the linguistic literature is to be believed, then options (a) and (c) convey crisp information, whereas (b) and (d) are vague (i.e., they permit borderline cases). From our own and previous experiments, there is no evidence that the fact in itself that an expression is vague should confer a 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 to the ones identified in our studies. These questions might focus on the amount of information that a given expression conveys (i.e., on granularity), on the avoidance of numbers, and on the use of evaluative terms. Let's see how this might pan out for the examples at hand.

First, the experiments by Mishra et al. suggest that it is important how much information is conveyed by an expression, and their findings are echoed by our own

thoughts about range reduction. In the case of (a)–(d) above, it appears that (a) conveys the most detailed information (designating the smallest segment of the temperature scale), followed by (b). Expression (b) appears to be followed by (d), and (d) by (c) (e.g., 40 degrees is above 25, but at 40 Celsius the word “warm" may no longer be applicable, giving way to words like “hot"):

$$a < b < d < c$$

If these hunches are correct then an *expert* may to prefer expression (a), because it gives her the most detailed information as a basis for her decisions. On the other hand, expression (d) (“warm") is shorter than the other three and avoids the use of numbers; 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 “warm" (see section 1 and (Peters et al., 2009)); the factor of evaluation is especially important if the hearer is unfamiliar with the metric used (e.g., because, being American, they are more used to Fahrenheit), in which case “warm" is much clearer. These considerations suggest that *non-experts* might prefer expression (d) over all others. If results of this kind were to be confirmed by experiments, they would be applicable in practical NLG, though care would need to be exercised to take the purpose of the communication into account.[3] In the experiments reported in this paper, the purpose of the utterances investigated was always clear (e.g., instructing people on a weight loss program (Mishra et al., 2011); informing a patient’s choice of hospital (Peters et al., 2009); or selecting a dot array (in our own studies).

One way to see why vagueness (as defined in our Introduction) may not matter as much for NLG – and for human communication – as is often thought, is the following thought experiment. Suppose a group of speakers understand the word “warm" as vague,

[3]One reason why “warm" might be less preferred than these considerations suggest is that, in the weather context, some very frequently occurring numbers, such as 20 (degrees), could be relatively easy to process, analogous to the subitizable numbers of (Trick & Pylyshyn, 1994), see our General Methodology section.

agreeing that temperatures above 26 count as warm, and temperatures below 24 do not count as warm, but considering temperatures between 24 and 26 as borderline cases. Now one day these speakers agree to sharpen up their definition deciding that, henceforth, “warm" means “> 25 degrees" (as in (c) above): this decision resolves the borderline cases, while everything else remains the same. It seems unlikely that this change in language use, from a vague meaning to a crisp one (i.e., one that has no borderline cases anymore), would lower the utility of the word. Our experimental findings are consistent with this idea.

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.

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.

Edgington, D. (1997). Vagueness by Degrees. In R. Keefe & P. Smith (Eds.), *Vagueness: a Reader. A Bradford Book.* The MIT Press, Cambridge, MA.

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

Fine, K. (1975). Vagueness, truth and logic. *Synthese*, *30*(3), 265–300.

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.

Green, M., & van Deemter, K. (2011). Vagueness as cost reduction: An empirical test. In *Proceedings of 'production of referring expressions' workshop at 33rd annual meeting of the cognitive science society.* Boston, MA.

Hobbs, J. R. (1985). Granularity. 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.

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.). (1997). *Vagueness: a Reader. A Bradford Book*. The MIT Press, Cambridge, MA.

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

Konstas, I., & Lapata, M. (2013). A global model for concept-to-text generation. *Journal of Artificial Intelligence Research*, *48*, 305–346.

Kuznetsova, A., Bruun Brockhoff, P., & Haubo Bojesen Christensen, R. (2016). 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` (Version 2.0-32)

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*, *22*(6), 733–738.

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.

Trick, L., & Pylyshyn, Z. (1994). Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychological Review*, *101*, 80–102.

Turner, R., Sripada, S., Reiter, E., & Davy, I. (2006). Generating spatio-temporal descriptions in pollen forecasts. In *Proceedings of EACL '06* (pp. 163–166). 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.* Springer-Verlag.

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.

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

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*(3), 338–353.