Running head: VAGUENESS AS COST REDUCTION

Vagueness as cost reduction

Matt Green (mjgreen@abdn.ac.uk) and

Kees van Deemter (k.vdeemter@abdn.ac.uk)

Department of Computing Science

University of Aberdeen

AB24 3UE

corresponding author: Matt Green

Rm 244, Computing Science

Meston Building

Aberdeen

AB24 3UE

email: mjgreen@abdn.ac.uk

tel: 07964339219

Abstract

Three experiments investigated the effects of vagueness in referring expressions, in order to test a prediction from game theory models of communication to the effect that crisp communications always have more utility than vague communications, under specified conditions. The experiments used a forced choice task, where participants were required to select an object on screen in response to an instruction in the form of a referring expression. The instruction referred to the number of dots the object contained, and was manipulated such that it referred either vaguely or crisply to the number. In Experiment 1, it was found that when either of two objects contained a subitizable (i.e., very small) number of dots, participants responded more quickly and accurately when the instruction was crisp than when it was vague. In contrast, in Experiment 2, using non-subitizable (i.e., larger) numbers, participants responded more quickly and accurately when the instruction was vague than when it was crisp, with diminishing returns for vagueness as the number of dots involved grew larger. Experiment 3 introduced an additional manipulation on whether the instruction contained a numeral, such that there were both vague and crisp versions of instructions that did and did not contain numerals. It was found that instructions that did not contain numerals were advantageous over those that did, and that within this effect, there was no reliable difference between vague and crisp alternatives.

Vagueness as cost reduction

Vagueness pervades the language that we use on a daily basis. In everyday use, language use may be called vague for various reasons.[1] In most academic use though, the word ‘vagueness’ has a more particular 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) (a similar definition can be found in Egre and Klinedinst (2011), among others). The most crucial of these criteria 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 from 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, which are used so frequently (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. Lipman argues that, 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 of communication that communicates the same state of affairs. The relevant conditions are, broadly speaking, that both interlocutors know all the relevant facts (e.g., both know the person's height precisely) and that the setting for the communication is co-operative. These conditions exclude such communicative situations as deliberate deception, and rhetorical situations like political debate and advertising where the intention is to persuade the interlocutor to adopt some point of view, where the persuasion might not be intended to be in the addressee's best interest.

Lipman observed that such 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, and how they might find a place in a game-theoretical explanation. In this paper, we focus on the first part of Lipman's question.

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 somehow 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 characterisation of the utility of vague language as the *cost reduction* hypothesis. The idea of vagueness as cost reduction can take various shapes (van Deemter, 2009), but the basic hypothesis is that it is easier for people to think in terms of loosely defined categories (such as "quite a few", or "many") than in terms of crisply defined ones (such as "thirteen", or "237"). This predicts that whenever a vague expression can perform the same communicative task as a precise one, it is rational to choose the vague expression. The corollary for comprehension is that vague expressions should be understood more readily than precise expressions.

Questions concerning optimal language use have many practical applications. Natural Language Generation[2] (henceforth NLG) systems must make decisions between different formulations of the same information. For example, if a man's height is 6 foot 2 inches, this could be expressed as “187.96 metres”, “6 foot 2”, or “tall”, among other ways, and the NLG system must decide between these. The problem is particularly relevant for NLG systems that take numbers as input, as many do. In the context of an NLG system faced with a practical decision of this kind, Lipman’s question becomes “Under what circumstances should vague terms be produced?” Relevant applications include weather forecasting on the basis of 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). At present, such NLG systems make decisions concerning the level of precision in the utterances that they generate (e.g., “the temperature will be in the high twenties tomorrow”) on the basis of little more than intuition. A better understanding of the benefits of different precision levels for readers would allow these systems to become more useful.

The cost reduction hypothesis is of direct relevance to psycholinguists interested in language comprehension, and additionally to psycholinguists interested in language production, for example in connection with the question of audience design (Clark & Murphy, 1982). For, to the extent that speakers and writers choose vague expressions over and above crisp ones because the former are easier to process for hearers than the latter, the cost reduction hypothesis suggests that speakers design their utterance for optimal benefit to their hearers – out of altruism, so to speak.

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 be used to illustrate both issues.

In a study of behaviour modification, Mishra, Mishra, and Shiv (2011) manipulated the presentation format of information about quantities. They compared information presented as a range (*between 0.5 and 1.5 g of cocoa*) with information presented as a single value at the midpoint of the range (e.g., *1 g of cocoa*). Information presented as a range was considered to be vague. The authors managed to effect behaviour modification in their participants, in the domains of mental acuity, physical strength, and weight loss, in laboratory and quasi-field studies. Their *vague* conditions improved performance relative to a *precise* condition. They explain the improvement in the range-of-values conditions as resulting from the participants' freedom to distort the information by focussing on one end of the range – the end that indicates proximity to the goal for motivated participants. The "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 midpoint-value 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 and Smith (1996) Egre and 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 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 a greater use of numerical information when making decisions; and greater use of an important indicator that is hard to interpret (i.e., survival rates). Participants' decisions were evaluated with respect to a 'gold-standard' set of interpretations derived from experts. It is interesting to consider whether their evaluative categories manipulation can be considered a manipulation of vagueness. Certainly, terms like *Fair* admit the possibility of borderline cases. However, when the visual boundary lines are taken into account, which map the terms to exact ranges, 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)*.[3] Accordingly, rather than the vagueness of categories such as *Poor*, Peters et al. emphasise the evaluative content inherent in these categories, and their affective potential.

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.

Game theory and the cost reduction hypothesis both make claims about vagueness in terms of its effects on people engaged in communicative linguistic acts. For game theory, when vagueness influences communication, the term *utility* is used to capture this influence: vagueness is said to have less utility than crispness. The cost reduction hypothesis notes that we use vague language frequently, and assumes that the reason for the high frequency of use is that vagueness brings about an *advantage*, for producer or comprehender. In this paper, we set out to measure these effects so that we might have some empirical basis for preferring one account over the other. An obstacle to this effort is that whereas we can obtain, in the laboratory setting, several quantitative measures of the effects of language, the opposing game theory and cost reduction accounts do not specify a metric in which *utility* or *advantage* should be measured. Therefore it is for us as experimenters to suggest a way to capture and measure the effects of vague and crisp expressions such that they can meaningfully be compared. This requires us to choose, and motivate the choice of, both a task that brings about some measurable behaviour (or alternatively, some measurable physiological state) that is consequent upon whether the language used to bring it about was vague or crisp; and also a measure of the behaviour that is sufficiently fine-grained to distinguish the vague case from the crisp case.

A good candidate for the task comes from the *forced choice* paradigm in which a participant is required to make a choice quickly among alternatives, indicating the choice by a physical response such as a button-press. These alternatives can be presented using vague and crisp language in such a way that the participant's choice is at least partly consequent upon whether the presentation was vague or crisp: indeed, in the case where other factors are held constant, it is plausible, and often assumed, that differences in the response are consequent to a great extent upon the mode of presentation. The button-press behaviour offers two metrics which are commonly taken to index cognitive load: response time and response accuracy, where fast accurate responses indicate low cognitive load, and slow or inaccurate responses indicate high cognitive load. In the case that vague presentations lead to lower cognitive load than crisp presentations, or vice versa, (where load is indexed by response time and accuracy), it is plausible to argue that this constitutes an advantage, or greater utility, of one presentation mode over the other.

In the experiments that we present here, we induced people to make a particular choice among alternatives by indicating one alternative using a referring expression. In the particular situation that we used, participants were presented with a number of objects on screen that constituted the alternatives. These objects were squares containing a number of dots, where the number in each was different. The referring expression referred to one square by indicating the number of dots it contained, and the participant was required to indicate that square by pressing the appropriate button on the keyboard. The referring expression offered a way to manipulate whether the square was indicated with a crisp or a vague expression of quantity, while holding other factors relatively constant. For example, the participant could be instructed "Choose the square with many dots" in the vague condition, or "Choose the square with 20 dots" in the crisp condition.

Experiment 1

In our first experiment we set out to compare responses to vague and crisp instructions in a forced choice task using response time and accuracy to measure the responses, and a variety of combinations of numbers. We aimed to identify cases where a difference was elicited such that these cases could form the focus of subsequent experiments.

Method

Participants. Twenty-five students and staff from the University of Aberdeen participated in the study in return for a cash payment of five pounds. Their median age was 23, ranging from 18 to 40. All participants self-reported fluency in English, and had normal (or corrected to normal) vision. Participants were recruited by advertising for participants on a university message board and a university mailing list, offering five pounds to volunteers.

Apparatus. A MacBook Pro laptop computer with a 13 inch screen presented the stimuli to the participants. Stimuli were created and presented using the language GNU

Octave (Eaton, 2002) and the Psychophysics Toolbox extensions (Brainard, 1997; Kleiner et al., 2007).

Design. We presented participants with 64 trials in experiment one in a single block. The basic properties of a trial were that it consisted of two elements: (1) two squares each containing a diffferent number of dots, and (2) a referring expression that referred to one of the squares by indicating the number of dots that it contained, leaving the other as a distractor. We measured two dependent variables: response time, and response accuracy. It was always the case that two squares appeared on the screen, each containing a number of dots: therefore items can be described using the notation $pair_{(p=1 \ldots 8)}([n], m)$ to indicate that for the stimulus pair p, one of the numbers of dots was n and the other m; and that the square-bracketed element $[n]$ was the target. There were 8 such pairs : $pair_{(1)}(2,4)$; $pair_{(2)}(3,5)$; $pair_{(3)}(2,6)$; $pair_{(4)}(3,7)$; $pair_{(5)}(6,8)$; $pair_{(6)}(7,9)$; $pair_{(7)}(4,8)$; $pair_{(8)}(5,9)$. See Table 1 for the properties of the stimuli.

Insert Table 1 about here

We manipulated four independent variables, and explicitly controlled one other, as follows.

The first independent variable was whether the item contained a subitizable number of dots. This was manipulated between items, and divided the items into two sets: one set comprised items that did contain a subitizable number; the other set comprised items that did not contain a subitizable number. There is evidence (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. Four of our pairs contained such a subitizable number, and four did not. This allowed us to ask whether vagueness or crispness might exert differential effects depending on whether this subitizing mechanism was, or was not, involved. The four pairs containing a subitizable number were: $pair_{(1)}(2,4)$; $pair_{(2)}(3,5)$; $pair_{(3)}(2,6)$; $pair_{(4)}(3,7)$; and the four pairs not containing a subitizable number were: $pair_{(5)}(6,8)$; $pair_{(6)}(7,9)$; $pair_{(7)}(4,8)$; $pair_{(8)}(5,9)$.

The second independent variable was whether the gap between the smaller and the larger number was small or large. At each level of the subitizable variable, half the items had a small gap, and the other half had a large gap. 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 numerosity of two sets of dots (Oeffelen & Vos, 1982) and for processing Arabic numerals and number words (Dehaene, 1996). By systematically manipulating this difference, we were able to ask whether effects of vagueness or crispness might differ according to whether the pair was a more or less discriminable pair.

The third independent variable was vagueness with two levels, vague and crisp: the effects of vague and crisp language were to be elicited by the way in which an instruction referred to an expression of quantity. Each stimulus was presented equally often with a vague expression of quantity, and a crisp expression of quantity. Taking the item $pair_{(1)}([2],4)$ as an example, where the target was the square with 2 dots, the crisp referring expression was "Choose the square with two dots" and the vague referring expression was "Choose the square with few dots".

Because we presented each item with the smaller number as target, and with the larger number as target, we were able to ask whether responses differed as a function of whether the target was large or small (with respect to the distractor). This constituted our fourth independent variable.

We also had to control which side the smaller number appeared on. This was manipulated within items such that each item had a version where the smaller number was on the left and a version where the smaller number was on the right. 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). Therefore we needed to present $pair_{(1)}(2, 4)$ once as $pair_{(1)}([2], 4)$ and once as $pair_{(1)}(4, [2])$. This accounts for two of the four presentations. The other two presentations were $pair_{(1)}(2, [4])$ and $pair_{(1)}([4], 2)$: i.e., with the larger number [4] rather than the smaller number as the target square; again once on the left and once on the right. This accounts for the eight presentations of each $pair_{(p)}(n, m)$.

Stimuli. Each stimulus consisted of two elements: (1) two squares containing dots; (2) the text of the referring expression indicating the target. Presentation of the two elements was simultaneous. The manipulation of vagueness was effected in the instruction, which referred to the number of dots in the target square as either (a) a cardinal numeral in the crisp conditions (i.e., *two, three, four, five, six, seven, eight, nine*), or (b) a linguistic quantifier in the vague conditions (e.g., *few, many*) . The position of the dots in each square was randomised per trial. An example stimulus is given in Fig. (1).

Insert Figure 1 about here

Procedure. The experiment was conducted in a small quiet room. On arrival in the room, the participant was told that he or she would be presented with objects on screen, together with an instruction to choose one of them by pressing the appropriate key on the keyboard (which was left cloverleaf for a target on the left, and right cloverleaf for a target on the right). The participant was instructed to respond quickly while avoiding errors.

There were 4 practice trials. After the practice trials the participant was invited to ask any questions they had about the procedure, and then the experimenter left the cubicle before the experimental blocks began. The order in which trials were presented was randomised per participant. Each trial would time out after 60 seconds if the participant did not respond. 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. Response time was measured from the appearance of the stimulus on screen until the button-press indicating the participant's choice. The trial sequence was keypress → instruction and stimulus → keypress.

Results

Preprocessing was different for the response time analysis and for the analysis of error rates. For the response time analysis, trials with response times measuring fewer than 250 ms were discarded, as were trials with response times of greater than 10,000 ms. Response times from trials with erroneous responses were also discarded. This led to the loss of 44 out of 1600 trials, which represented 2.75% of trials. One participant was excluded from the analysis of variance due to making errors in all responses in one condition. This would have led to an unbalanced design for ANOVA had the participant been retained. This participant's exclusion led to the loss of another 59 trials, the other 5 having been excluded by the time and error trimming described earlier in this section. Overall, 6.4% of trials were excluded from the response time analysis (103 of 1600 trials). For the analysis of error rates however, no trials were discarded.

A 4-way (2 x 2 x 2 x 2) repeated measures ANOVA was carried out on (correct response) RT as the dependent variable, with gap (small:2 or large:4), subitizability (whether the display included a suitizable number or not), quantity (whether the larger or smaller number was the target), and vagueness (vague instruction or crisp instruction) as independent variables each with two levels.

Grand mean RT was 1637 ms, with standard error 114 ms. The manipulations revealed significant main effects of gap size, target quantity, and subitizability, but no significant main effect of vagueness. There were reliable interactions of vagueness with quantity; and of gap with subitizability, and a reliable three-way interaction between vagueness, quantity, and subitizability. No other interactions were reliable. Details of the reliable effects and interactions follow.

Vague instructions led to numerically slightly longer RT than precise instructions but this difference was not significant (1600 ms vs 1673 ms, $F(1,23) = 2.1, p = .161, \eta^2 = .084$). Participants responded significantly faster when the numerical gap was large at 4 dots than when it was small at 2 dots (1397ms vs 1876 ms, $F(1,23) = 70.6, p < .001, \eta^2 = .754$, halfwidth of the 95% confidence interval for the 479 ms difference was 118 ms). Hereafter halfwidths are referred to simply as CI. When one of the squares contained a subitizable number of dots, participants responded significantly faster than when neither of the squares contained a subitizable number of dots (1292 ms vs 1981 ms, $F(1,23) = 83.9, p < .001, \eta^2 = .785$, 95% CI of the 689 ms difference was 155 ms).

When participants were instructed to choose the square with the larger number of dots, they responded 165 ms more slowly than when they were instructed to choose the square with the smaller number of dots, for the same pair of numbers (1719 vs 1554 ms, $F(1,23) = 15.4, p < .01, \eta^2 = .401,$95% CI of the 165 ms difference was 86.5 ms). Vagueness interacted significantly with quantity ($F(1,23) = 25.065, p < .001, \eta^2 = .521$), such that vagueness was a significant advantage of 116 ms when participants were instructed to choose the square with the larger number of dots (1761ms vs 1645 ms, $t(23) = 2.17, p < .05$, 95% CI of the 116 ms difference was 111 ms) but a significant disadvantage of 275ms when participants were instructed to choose the square with the smaller number of dots (1409 ms vs 1684 ms, $t(23) = -3.88, p < .001$, 95% CI of the 275 ms difference was 147 ms).

The effect of a subitizable number in one of the squares varied as a function of the size of the gap ($F(1, 23) = 49.669, p < .001, \eta^2 = .683$). In the big gap conditions, the advantage for a subitizable number was smaller than it was in the small gap conditions. In the big gap conditions, the advantage due to a subitizable number was 308 ms (1551 ms vs 1243ms, $t(23 = 7.5, p < .001$, 95% CI of the 308 ms difference was 85 ms), whereas in the small gap conditions, the advantage due to a subitizable number was greater at 1059 ms (2398 ms vs 1338 ms, $t(23) = 8.6, p < .001$, 95% CI of the 1059 difference was 254 ms).

There was a reliable three-way interaction between subitizability, quantity, and vagueness ($F(1, 23) = 6.653, p < .05, \eta^2 = .224$), such that vague instructions were advantageous when subjects were instructed to choose the bigger quantity and there was no subitizable number, but disadvantageous when the participants were instructed to choose the smaller number, regardless whether there was a subitizable number or not. This interaction is plotted in Fig. (2).

The grand mean participant error rate was low at 1.3%. A generalized linear mixed model (Jaeger, 2008) was fit to the error data, with task, quantity, vagueness and subitizability, as well as all interactions between them, as fixed effects, and with random intercepts for subject and item. None of the main effects or interactions showed reliable effects, apart from the main effect of gap. Mean error rates rose by 2.2% in the small gap conditions (mean error rate in the big gap conditions: 1.38%; mean error rate in the small gap conditions: 3.63%, $Wald\ Z = 2.323, p < .05$).

Insert Figure 2 about here

Discussion

Response times did not differ reliably according to whether the quantifier was vague or crisp, all other things being equal, but vagueness exerted effects in interactions with other effects.

Moyer and Landauer (1967) observed that participants compared two numbers faster as the numerical distance, or gap, between them grew larger. Our subjects showed a similar effect, responding nearly 500 ms faster (and with fewer than half as many errors) to the instruction to choose a square containing a number of dots when the numerical distance was four than when it was two. This effect did not interact with vagueness for response times or error rates, suggesting that participants respond similarly in the small and large gap conditions regardless whether the instruction is vague or precise.

Similarly in line with previous findings (e.g., Mandler & Shebo, 1982), our participants responded faster when one of the squares contained a subitizable number of dots than when neither did. Vagueness did not interact with this effect, suggesting that participants' rapid identification of very small quantities is not affected by whether they are identified using precise or vague quantifiers. The advantage for a subitizable number was smaller when the gap was big than when the gap was small. This interaction can be explained if a big gap size confers a nearly maximal benefit that is not improved by the presence of a subitizable number. When gap size is small, the task is harder, and the benefit conferred by a subitizable number has an opportunity to manifest itself.

Participants were faster to choose the square with the smaller number of dots than they were to choose the square with the larger number of dots. Vagueness exerted a considerable influence on this effect, acting like a damper to even out response speeds across the two conditions. A vague quantifier sped up the choice of the larger number of dots; and slowed down the choice of the smaller number of dots.

The results showed a three-way interaction between subitizability, quantity and vagueness, which is interpreted here. When participants were instructed to choose the smaller number of dots rather than the larger number, they responded faster to instructions that mentioned the number than to vague instructions that identified the same square, and although the presence of a subitizable number speeded responses, it did so equally for vague and numerical quantifiers. In contrast, when participants were instructed to choose the larger number of dots, participants responded slower to precise instructions than to vague instructions, but only when there was no subitizable number. Cases where vagueness was advantageous include $pair_{(6)}(7,9)$ with an instruction to choose the square with many dots; $pair_{(5)}(6,8)$ with an instruction to choose the square with many dots. Cases where vagueness was disadvantageous include $pair_{(1)}(2,4)$ with an instruction to choose the square with few dots; $pair_{(6)}(7,9)$ with an instruction to choose the square with few dots. This suggests that participants had to count the dots only when there was no subitizable number, or in other words that vague quantifiers confer a benefit over precise quantifiers only when the precise quantifiers identify a large number of dots.

Experiment 2

Experiment 1 showed that vagueness does not exert much influence on the processing of numbers in the subitizable range. Evidence suggests that such numbers are processed very quickly using a different system than larger numbers (e.g., Trick & Pylyshyn, 1994). For experiment two we wanted to avoid the potential confound introduced by the inclusion of numbers in the subitizable range. We expected that vagueness would be positively beneficial when larger numbers of dots were involved. This is because we observed reliable advantages of vagueness in the pilot experiment only for choosing the bigger of two numbers of dots (and then only when there was no subitizable number in the display). We decided to use larger numbers of dots in order to explore this finding more thoroughly. We used more levels of gap in this experiment. This meant that we had a kind of index of discriminability, with small gaps resulting in displays with low discriminability, and bigger gaps resulting in increasingly discriminable displays. Because we always used a baseline number of 25 dots, it was possible to directly compare different gap sizes against the same constant baseline.

We also made a change to the procedure of the experiment. In experiment one the response time that we used as a dependent variable included the reading of the instruction – the participant pressed the key to initiate the trial and the response time before reading the instruction. This could have conflated two separate processes: comprehending the instruction; and selecting a response. In experiment two we tried to separate these processes. We did this by requiring the participant to read the instruction before pressing the key that started the trial and displayed the squares and the dots. The response time was measured from this key press. In this way the response time for experiment two should index only the selection process.

We hypothesised that under these conditions: (a) RT will be faster for vague than precise instructions; (b) Vagueness will be more advantageous at smaller distances (i.e., low discriminability) than at large distances (i.e, high discriminability).

Method

Participants. Twenty participants were recruited and paid in the same way as in experiment one, and similarly aged between 18 and 45, with a median age of 26. All participants self-reported fluency in English, and had normal, or corrected-to-normal vision.

Apparatus. A MacBook Pro laptop computer with a 13 inch screen presented the stimuli to the participants. Stimuli were created and presented using the language GNU Octave (Eaton, 2002) and the Psychophysics Toolbox extensions (Brainard, 1997; Kleiner et al., 2007).

Design. We presented participants with 256 trials, arranged in 4 blocks of 64 trials each. It was always the case that two squares appeared on screen. One of these always contained 25 dots, and the other contained a number of dots that varied, and which was always the target. We can use the notation $pair_{(p)}(n, 25)$ to represent a pair. The pairs were: $pair_{(1)}(5, 25)$; $pair_{(2)}(10, 25)$; $pair_{(3)}(15, 25)$; $pair_{(4)}(20, 25)$; $pair_{(5)}(30, 25)$; $pair_{(6)}(35, 25)$; $pair_{(7)}(40, 25)$; $pair_{(8)}(45, 25)$.

We manipulated two independent variables: gap size, with four levels (gap size = 5, 10, 15, or 20), and vagueness, with two levels (crisp and vague). We had to control another two variables: target size (whether the referring expression indicated the larger number of dots or the smaller number of dots); and target side (for a given level of target size, whether the target appeared on the left or the right). We measured, as our dependent variables, response time and response accuracy.

Our manipulation on gap size divided the items into four groups, one for each gap size. Items with gap size 5 were: $pair_{(4)}(20, 25)$; and $pair_{(5)}(30, 25)$; items with gap size 10 were: $pair_{(3)}(15, 25)$; and $pair_{(6)}(35, 25)$; items with gap size 15 were $pair_{(2)}(10, 25)$ and $pair_{(7)}(40, 25)$; and items with gap size 20 were $pair_{(1)}(5, 25)$ and $pair_{(8)}(45, 25)$.

Our manipulation on vagueness meant generating crisp and precise versions of each referring expression. Each pair was presented 8 times with a vague expression of quantity, and 8 times with a crisp expression of quantity. Taking $pair_{(1)}(5, 25)$ as an example, the crisp referring expression was “Choose the square with 5 dots”, and the vague referring expression was “Choose the square with few dots”.

We needed to control which side the target was presented on. Each pair was presented equally often with with the target on the left, as with the target on the right.

We also needed to control whether the target was the smaller or larger number. The items with numbers less than 25 (pairs 1 to 4) formed a group with the smaller number as target and the other items (pairs 5 to 8) formed a balancing group with the larger number as the target number

Table (2) shows the orgainisation of trials in experiment 2.

Insert Table 2 about here

Stimuli. Each stimulus consisted of two elements (a) the text of the referring expression indicating the target; (b) two squares containing dots. Where a number was mentioned in the instruction, it was always in the form of an arabic numeral (e.g., 1,2). This contrasts with experiment one, where numbers were given in natural language form. The position of the dots was randomised per-trial. Fig (3) gives an example stimulus.

Insert Figure 3 about here

Procedure. The experiment was conducted in a quiet room. On arrival in the rooom, the participant was told that he or she would be presented with an instruction to choose one of two squares by reference to how many dots it contained. Participants were required to press the space key after reading the instruction. Then there was a central fixation cross for 1000 ms, and a blank screen for 500 ms, followed by the squares and dots (without repetition of the referring expression). Response time was measured as the latency between the presentation of the dots and squares, and the keypress identifying the decision. The display would stay on screen until the participant responded (there was no timing-out). Participants were asked to respond quickly while avoiding errors. There were 8 practice trials, after which the participant was invited to ask any questions about procedure. After answering these the experimenter left the cubicle for the duration of the experiment. The order in which trials were presented was randomised per-participant, and stimuli were presented in 4 blocks of 64 trials each, between which the participant could rest. 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

Because the numbers of dots in this experiment were much greater than in experiment one, no maximum RT cutoff was imposed. The same minimum cutoff as experiment one, 250 ms, was used, but no trial had RT less than 250 ms in this experiment. RTs for trials with erroneous responses were discarded, leading to the loss of 354 trials from 5120, representing 6.9% of the trials.

Means for response times and error rates are given in Fig. (4).

An omnibus 2-way (2 x 4; vagueness x gap size) ANOVA was carried out on participant mean correct RT as the dependent variable, and Vagueness and Gap size as the independent variables manipulated within subjects. This analysis averaged RT over Quantity. Grand mean RT was 1088 ms with standard error 125 ms, 95% confidence interval from 826 ms to 1350 ms. Vague quantifiers attracted reliably faster mean RT than precise quantifiers (precise: 1246ms, vague: 930 ms, $F(1,19) = 12.106, p < .01, \eta^2 = .389$). There were reliable differences across RT at the different levels of gap (main effect of gap: $F(1.2, 22.793) = 17.839, p < .001, \eta^2 = .484$, after Huynh-Feldt correction for sphericity violation). Gap and vagueness interacted reliably ($F(1.711, 32.5 = 7.498, p < .01, \eta^2 = .283$, after Huynh-Feldt correction for sphericity violation). RTs reliably grew faster as the gap size increased (overall linear trend: $F(1,19) = 20.127, p < .001, \eta^2 = .514$). The linear trend was significant at each level of vagueness (linear trend in the precise conditions: $F(1,19) = 23.611$; linear trend in the vague conditions: ($F(1,19) = 10.946, p < .01$)). This linear increase in speed differed reliably between the precise and vague conditions ($F(1,19) = 14.140, p < .01, \eta^2 = .427$).

In the precise conditions RTs started out much slower than in the vague conditions, at the smallest gap size, but the two conditions converged to very fast times at the largest gap size. The interaction of the linear trend across gap size with vagueness is plotted in Fig. (4).

A generalized linear mixed model (Jaeger, 2008) was fit to the error data, with vagueness and gap and their interaction as fixed effects, and with subject as a random effect. Vagueness exerted a reliable effect on accuracy, with precise instructions reliably inducing more errors than vague instructions: $Wald\ z = 2.326, p < .05$). There was a reliable linear trend in error rates across gap, with error rates falling as gap size increased $(Wald\ z = -10.203, p < .001)$. This linear trend interacted with vagueness $(Wald\ z = -5.366, p < .001)$. Error rates were greater in the precise conditions than the vague conditions when gap size was small, and that this difference diminished with increasing gap size until it reversed at the biggest gap size.

Insert Figure 4 about here

Discussion

The experiment provided support for hypothesis 1, that RT would be faster for the vague instruction conditions than for the precise instruction conditions. The experiment also provided support for hypothesis 2, that vagueness would be more advantageous at smaller than at larger distances. The findings of experiment two can be summarised thus: vagueness is easier to process than crispness, and returns for vagueness diminish as the stimuli become more easily discriminable. We consider here four explanations of this pattern.

(a) 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 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. A possible explanation for the pattern of results is cast in terms of the precision of an estimate of quantity that is required to carry out the task. For example, in the display $\text{pair}_{(1)}(5, 25)$, with a large distance, vague instructions (*many*, *few*) might be able to be carried out by a visual comparison that identifies one square as more numerous than the other. Precise instructions (to identify the square with 5 dots) might be able to be carried accurately out with an estimate with a fairly large error of 19. In the display $\text{pair}_{(1)}(5, 25)$, which has a small distance, vague instructions might be able to be be carried out with a visual comparison that identifies one square as more numerous than the other. Precise instructions might require an estimate with a maximum error of 4 in this case. The pattern of results can be explained if we assume that the degree to which an estimate with error 4 is harder than an estimate with error 19 is greater than the degree to which a visual comparison between 20 and 25 is harder than a visual comparison between 5 and 25.

(b) An account that we will dub the *instruction format hypothesis*, or the number / no-number account, explains the vagueness advantage by placing importance on the observation that the vague conditions did not mention a numeral whereas the crisp conditions did. This could make the difference between a task that taps numerical processing, and one that taps magnitude judgements. The diminishing returns for vagueness can be explained by assuming that the numerical task is particularly challenging when the stimuli are less-discriminable, whereas the magnitude judgement is relatively less difficult in the less-discriminable trials.

(c) Rayna and Brainerd (1989) can account for the pattern in a framework they called the *hierarchy of gist*. They proposed that reasoners encode representations at different levels of precision, and that these representations can be ordered with respect to precision, forming a hierarchy of gist. They further claimed that reasoning operated at the lowest level in this hierarchy that would allow one to accomplish the assigned task. Among the levels of this hierarchy are (1) ratio representations at the top, which encode numerical information exactly; (2) ordinal representations that capture relative magnitude but do not encode numerical differences, i.e., they establish a rank ordering of quantities by magnitude, e.g., *large*, *larger*, *largest*; (3) nominal or categorical representations that capture only the presence of absence of quantity e.g., *some lives*; *no lives*. Brainerd and Gordon (1994) presented evidence that children spontaneously encode relative numerosity, at the ordinal level of the hierarchy mentioned above, when they are presented with stimuli that vary in numerosity. The hierarchy of gist offers an explanation of our vagueness advantage in experiment two that is independent of vagueness in the Keefe and Smith (1996) sense. In experiment two, the participants were presented with the instruction for the trial first, and then with the squares and dots. This was done in order to try to separate the processes involved in processing the instruction from those involved in processing the stimuli. A consequence of this serial presentation though is that participants can infer from the instruction which level of representation is the minimum level necessary to carry out the impending task successfully. We do not suggest that this takes the form of a conscious deliberation – rather it might be apprehended spontaneously in the same way that the children in Brainerd and Gordon (1994) did without conscious deliberation. For example, it follows naturally from the instruction *Please choose the square with fewest dots* that an ordinal representation will be necessary and sufficient for the purposes of the current trial: similarly it follows naturally from the instruction *Please choose the square with 34 dots* that a higher level of the hierarchy will be necessary and suficient for the purposes of this trial - the ratio level that encodes numerosity exactly. We found a response time advantage in experiment two for instructions that used the quantifiers *many* and *few*, when compared with times for quantifiers like *15* and *45.* In the terms of Rayna and Brainerd (1989), this can be explained as an advantage for the sufficiency of a lower level of representaion – without reference to vagueness.

(d) The *two-systems account* of decision making (e.g., Sloman, 2007) can offer an explanation of the pattern. Dual process models of decision making propose two systems of processing: the quick and affective System 1 and the deliberative and rule-based System 2. This two-systems account can explain our finding of diminishing returns for vagueness in experiment 2. When the gap is small, participants use System 2 (slow, deliberative) for the precise instructions. When the gap is large, they take a shortcut for the precise conditions and merely establish which square is more (or less) numerous, using System 1 (quick, heuristic). In the vague conditions, they use the heuristic system whether the gap is small or large. This account explains the diminishing returns for vagueness that we observed in terms of a change of strategy in the baseline precise conditions as the gap size grew larger.

The potential for vagueness in the vague conditions can be said to have been unrealised in experiment two. In both experiment one and experiment two, the vague instructions were *choose the square with many dots* and *choose the square with few dots. Many* and *few* have the potential for vagueness defined as the existence of borderline cases. Given a square with 15 dots, and asked the question 'are there many dots in the square?', we can imagine that some people would be unsure (contrast this with the questions 'are there fewer than 20 dots in the square?' to which there is no room for uncertainty). However, this potential for vagueness may be unrealised in the context of a choice between two squares. Given two squares, one with 20 dots, and another with 25 dots, and asked to 'choose the square with few dots', it seems that no one could be unsure which square is the intended target. If there is no room for uncertainty, then the potential for *many* and *few* to be vague can be said to be unrealised. On these grounds experiments one and two can be said not to have used vague conditions with true vagueness.

Experiment 3

In experiments one and two, the participants chose one of two squares. 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 the first two experiments, the quantifiers in the vague conditions did not really meet this definition. This is because there were no borderline cases of the referent that could make the referent set 'not well-defined'. Experiment three used three squares so that the vague quantifiers always had more than one possible referent. To enhance the potential for the vague quantifiers to have true vagueness, we also used indefinite articles in the vague instructions. Thus, instead of the instruction *choose the square with few dots*, we used *choose a square with few dots*.

The core finding in experiment two was that vagueness exerted a beneficial effect compared to precise equivalents in the domain of quantifiers. A problem with this finding is that vagueness in the strict sense that we are interested in was confounded in that experiment with the format of the quantifier. For example, contrast an expression from the vague condition: 'the square with few dots' with an expression from the precise condition: 'the square with 15 dots'. One difference is that 'few' is vague (or at least has the potential for vagueness) and '15' is precise. Another difference is that 'few' is a linguistic, or verbal quantifier while '15' is a numerical quantifier, in the sense that a number is mentioned explicitly. Since these two differences were confounded in experiment two, the finding is vulnerable to an alterative interpretation, that our vagueness advantage was in contrast an advantage for the linguistic or verbal form of the quantifier. The present experiment, experiment three, pitted these two alternative interpretations against each other in a factorial design.

The cost reduction hypothesis predicts that there will be a main effect of vagueness such that the vague instructions confer a processing advantage relative to the precise instructions. In contrast, the instruction format hypothesis predicts that there will be a main effect of instruction format, with verbal format being advantageous compared to numerical format.

Method

Participants. Thirty participants were recruited and paid in the same way as experiment 2. They were aged between 18 and 45 with a median age of 28. All participants self-reported fluency in English, and had normal, or corrected-to-normal vision.

Apparatus. A MacBook Pro laptop computer with a 13 inch screen presented the stimuli to the participants. Stimuli were created and presented using the language GNU Octave (Eaton, 2002) and the Psychophysics Toolbox extensions (Brainard, 1997; Kleiner et al., 2007).

Design. We presented participants with 256 trials, arranged in 4 blocks each with 64 trials.

We used four different triples of numbers: $triple_{(1)}(6, 15, 24)$; $triple_{(2)}(16, 25, 34)$; $triple_{(3)}(26, 35, 44)$; $triple_{(4)}(36, 45, 54)$. Each triple had the following properties: it comprised three squares (instead of two as in experiments one and two); the central number was always presented in the middle of the three; there were two flanking numbers where one was smaller than the central number and one was bigger. The triples are given in Table 3.

Insert Table 3 about here

We manipulated two independent variables: Vagueness with two levels (vague and crisp); and Format with two levels (numerical, verbal). This yielded four conditions (vague numerical; vague verbal; crisp numerical; crisp verbal). Each condition had a different referring expression, as follows, using $triple_{(1)}(6, 15, 24)$ as an example. The vague numerical condition's referring expression was "Choose a square with about 10 dots". No square contained 10 dots. 10 is slightly closer to 6 than to 15. Therefore the best referent for this referring expression was the square with 6 dots; the borderline response was the square with 15 dots; and the poorest referent was the square with 24 dots. The crisp numerical conditions's referring expression was "Choose the square with 6 dots" on half the presentations in that condition and "Choose the square with 24 dots" on the other half. One square did contain the exact number mentioned. The square with 15 dots was the borderline response; and the remaining square was the poorest referent. The vague verbal condition's referring expression was "Choose a square with few dots" on half of the presentations, and "Choose a square with many dots" on the other half. For this condition, the best referent was the square with 6 dots for 'few' and 24 dots for 'many'; the borderline case was the square with 15 dots; and the remaining square was the poorest referent. The vague numerical condition's referring expression was "Choose the square with the fewest dots" on half of presentations, and "Choose the square with the most dots" on the other half. For this condition, the square with 6 dots (for few; 24 dots for many) was the best referent; the square with 15 dots was the borderline case, and the remaining square was the poorest referent. Table 4 gives examples of an instruction from each condition.

Insert Table 4 about here

Each triple was presented in each condition 16 times. 8 of these identified the larger number, and 8 the smaller.

We measured two dependent variables: response time; and the probability of a participant choosing the borderline case.

Stimuli. Fig (5) gives an example stimulus. First, the referring expression that constituted the instruction for that trial was displayed. The participant then pressed a key to indicate that he or she had read the instruction. After 1 second, the squares and dots were presented, while preserving the text of the referring expression. The position of the dots in the squares was randomised per-trial.

Insert Figure 5 about here

Procedure. The experiment was conducted in a small quiet room. On arrival in the cubicle, the participant was told that he or she would be presented with objects on screen and required to choose one in response to an instruction on screen, by pressing the button corresponding with the object. There were 5 practice trials. After the practice trials the experimenter left the room. There were 4 blocks of 64 trials each. In between blocks the participant had the opportunity to rest before continuing. The response time dependent variable was measured from the presentation of the squares and dots, until the keypress indicating the participant's choice. The trial would timeout after 60 seconds if there was no response. The dependent variable measuring whether the participant chose the borderline case was also recorded at this time. 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 borderline case or not, and how long it took the participant to respond.

Results

No responses were treated as erroneous, because errors were essentially undefined for the vague instructions (e.g., 'about ten'). It was noted whether participants chose the borderline square. Trials with RT less than 250 ms, and trials with RT greater than 50,000 ms were discarded from RT analysis. This led to 3 trials being discarded, representing less than one percent of the datapoints (.039% of the datapoints; 3 of 7680 trials).

A three-way (2 x 2 x 2) repeated measures ANOVA was carried out, with Quantity (2 levels, instruction to choose the biggest number; instruction to choose the smallest number), Task (2 levels, Numerical instruction, Verbal instruction), and Vagueness (2 levels, Crisp and Vague) entered as independent variables manipulated within-subjects in a factorial design. Grand mean RT was 2842 ms, with standard error 331 ms. There was a significant main effect of task, but no other main effects were significant, and no interactions were significant. Figure (6) shows response time as a function of stimulus and instruction format. Mean responses to the instructions to choose the biggest square were slightly faster than to the instructions to choose the smallest square, but the difference was not significant (2817 vs 2868 ms, $F(1, 29) < 1$). Responses to the numerical task instructions were reliably much slower than responses to the verbal task instructions (3773 ms vs 1911 ms, $F(1, 29) = 13.369, p < .01, \eta^2 = .316$). Mean responses to the crisp instructions were slightly faster than to the vague responses, but the difference only approached significance at the 5% level (2751 vs 2933 ms, $F(1, 29) = 3.470, p = .073$). The quantity x vagueness interaction was not reliable ($F(1, 29) = 2.913, p = .099$). The task x vagueness interaction was not reliable ($F(1, 29) = 2.311, p = .139$). The quantity x task interaction was not reliable ($F(1, 29) = 1.785, p = .192$). The three-way interaction between quantity, task, and vagueness was not reliable ($F(1, 29) = 1.753, p = .196$).

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 quantity as fixed effects, and with random intercepts for subject and item. Results are plotted in Figure (6). There was a significant main effect of each of task, vagueness, and quantity; a significant interaction between task and vagueness; and a significant three-way interaction between task, vagueness, and quantity. Participants were significantly more likely to choose the borderline option for vague instructions than for precise instructions (21.9% vs 11.3%, $p < .001$), and reliably more likely to choose the borderline response for the smaller quantity than for the larger quantity (12.2% vs 20.9%, $p < .001$). Participants were also reliably more likely to choose the borderline square when the instruction used the numerical format rather than the verbal format (30.1% vs 3.0%, $p < .001$). Task and vagueness interacted reliably ($p < .001$) such that the propensity to choose the borderline case in response to numerical format instructions was greater when they were expressed vaguely (e.g., ‘about 20’). The three-way interaction between task, vagueness, and quantity was due to bigger quantity generally being less likely to result in a borderline response, except in the verbal, vague conditions where instructions to choose the bigger quantity were more likely to result in a borderline response.

Insert Figure 6 about here

Discussion

This experiment was designed to test the hypothesis that vagueness would interact with instruction format. If found, such an interaction would show that the benefit for vagueness that was observed in experiment two was additional to any benefit due to the verbal instruction format. There was no such interaction. Instead, instruction format exerted a large effect that was not reliably modulated by vagueness. Thus it appears that the vagueness advantage elicited in experiment two was really a benefit of the verbal instruction format over the numerical format.

A potential confound is that the no-number quantifiers are asymmetrical while the precise quantifiers are symmetrical. For example, given triple$_{(1)}(6, 15, 24)$, the expressions *about 10* and *10* are symmetric about 10, while the expressions *few* and *fewest* are asymmetric about 10 – the distribution is skewed towards lower numbers. Support for this confound is that in the skewed conditions, the borderline case was chosen less frequently than in the symmetric conditions. In other words, people were more likely to choose the middle value when the distribution encompassed this value than when it was skewed away from it.

In experiment three, where the potential for vagueness was realised in the form of borderline cases, verbal-vague (the equivalent of the vague condition in experiment two) attracts faster times than numerical-crisp (the equivalent of the crisp condition in experiment two). This is the same pattern that we observed in experiment two. This indicates that the vague conditions in experiment two were not faster as a result of the lack of borderline cases; but rather because they did not specify a number in the instruction.

General Discussion

Several answers to Lipman's question – why vagueness is as prevalent as it is in human communication – have been suggested (van Deemter, 2009, 2010). In this paper, we have tested one them, namely the cost reduction hypothesis. As far as we know, no other answers to Lipman's question have been subjected to experimental scrutiny yet. Has the cost reduction hypothesis been confirmed by our experiments? In other words, do readers benefit from vagueness, as the cost reduction hypothesis asserts?

At one level, the answer to these questions is affirmative. In all our experiments, we have found consistent and fairly substantial benefits, in terms of hearers' reaction times, for referring expressions involving vague quantifiers like “many” over crisp ones involving numerical quantifiers, such as “25”. An example of an NLG domain that could benefit from these results is the GIVE challenge (Byron et al., 2009) (Koller et al., 2010). This is an environment where NLG systems issue instructions to human comprehenders who are navigating a virtual environment. A typical situation might involve a panel with ten buttons, for example, that the human should approach in the virtual world in order to press a particular button, when more than one panel with buttons on it is available in the immediate environment. Our results suggest that an instruction like *approach the panel with many buttons* that refers to the panel without reference to the particular number might be processed faster by the human player than an instruction that specifies the number like *approach the panel with ten buttons*, and therefore be a better referring expression for the NLG system to produce. This is a potentially useful result, although it is as yet unclear whether it generalises to situations where the quantifier is not used as part of a reference task.

At a deeper level, the answer must be negative, because our third experiment shows that the benefits reported in the previous paragraph were not a consequence of the vagueness of the words involved (i.e., of the fact that they allow borderline cases), but of the fact that they used a non-numerical expression, that is, the description does not rely on counting. In other words, if the choice for the NLG system mentioned above is broadened to include expressions that are both verbal and crisp, as well as ones that are both numerical and vague, then the only advice that can be given on the basis of our experiments is that the numerical ones should be avoided; we found no clear difference, in terms of benefit for readers, between verbal descriptions that are crisp and those that are vague.

References

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

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

Brainerd, C., & Gordon, L. (1994). Development of verbatim and gist memory for numbers. *Developmental Psychology*, *30*(2), 163.

Byron, D., Koller, A., Striegnitz, K., Cassell, J., Dale, R., Moore, J., et al. (2009). Report on the First NLG Challenge on Generating Instructions in Virtual Environments (GIVE). In *Proceedings of the 12th European Workshop on Natural Language Generation* (pp. 165–173).

Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. In J. Le Ny & W. Kintsch (Eds.), *Language and comprehension* (Vol. 9, p. 287 - 299). North-Holland.

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.

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

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*.

Koller, A., Striegnitz, K., Gargett, A., Byron, D., Cassell, J., Dale, R., et al. (2010). Report on the second NLG challenge on generating instructions in virtual environments (GIVE-2). In *Proceedings of the 6th International Natural Language Generation Conference* (pp. 243–250).

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`)

Mandler, G., & Shebo, B. (1982). Subitizing: An analysis of its component processes.

Journal of Experimental Psychology: General, *111*(1), 1–22.

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

Moyer, R., & Landauer, T. (1967). Time required for judgements of numerical inequality. *Nature*, *215*, 1519–1520.

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

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., et al. (2009). Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, *173*(7-8), 789–816.

Rayna, V. F., & Brainerd, C. J. (1989). Fuzzy-trace theory of framing effects in choice. In *Proceedings of the 30th Annual Meeting of the Psychonomic Society.* Atlanta, GA.

Reiter, E., & Dale, R. (2000). Building natural language generation systems.

Sloman, S. (2007). Two systems of reasoning. In T. Gilovich, D. Griffin, & D.Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment.* CUP.

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 *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.

Footnotes

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

[2] NLG systems take data or formulas as input, and transform them into natural language outputs (Reiter & Dale, 2000). The process parallels language production in humans.

[3] This distinction between the potential for vagueness, and the realisation of vagueness in the context of a particular stimulus, is something we will return to when discussing our own experiments.