Generating Quantified Referring Expressions with Perceptual Cost Pruning

Gordon Briggs[1] and **Hillary Harner**[2]
[1]Navy Center for Applied Research in Artificial Intelligence
[2]NRC Postdoctoral Fellow
U.S. Naval Research Laboratory, Washington, DC 20375 USA
{gordon.briggs,hillary.harner.ctr}@nrl.navy.mil

Abstract

We model the production of quantified referring expressions (QREs) that identify collections of visual items. To address this task, we propose a method of perceptual cost pruning, which consists of two steps: (1) determine what subset of quantity information can be perceived given a time limit t, and (2) apply a preference order based REG algorithm, such as the Incremental Algorithm (IA), to this reduced set of information. We demonstrate that this method successfully improves the human-likeness of the IA in the QRE generation task by successfully modeling human-generated language in most cases.

1 Introduction

Production of natural and human-like referring expressions in visual contexts is an ongoing challenge in natural language generation (NLG). What makes referring expression generation (REG) in visual contexts difficult is that it strongly depends on the dynamics of human perception (Clarke et al., 2013; Elsner et al., 2018). How to integrate REG algorithms with dynamic and incremental human-like perception is still an open problem. Starting with the Incremental Algorithm (IA) (Dale and Reiter, 1995), REG algorithms have sought to model factors of perceptual salience by considering a preferred ordering of visual attributes. Building off of the IA, the Visual Object Algorithm (VOA) favors certain visual attributes based on relative perceptual cost (Mitchell et al., 2013). However, these algorithms assume complete knowledge and model perceptual cost through preference orderings. In practice, some visual information is not just dis-preferred, but *impossible* to ascertain under time constraints.

In this paper, we investigate a more radical means to integrate perceptual cost in REG: pruning a knowledge base according to perceptual costs. Specifically, we propose using only information in a preference-based REG algorithm (i.e., IA) that a psychological model suggests is plausible to have encoded under particular time constraints. To evaluate this proposed solution, we examine the problem of quantified referring expression (QRE) generation. First, we introduce QREs and give a brief overview of the psychology of numerical perception. Second, we detail a mathematical model of the time requirements of exact numerical perception and how we apply the proposed method of perceptual cost pruning (PCP) to the QRE task. We demonstrate that this method successfully improves the human-likeness of the IA in the quantified REG task by successfully modeling human-generated language in most cases.

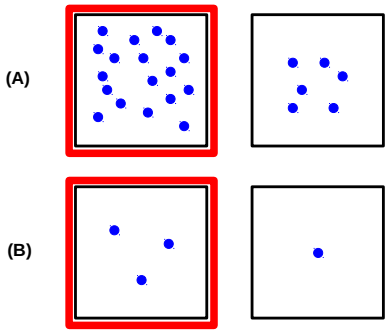

Figure 1: Examples of QRE generation tasks. Target referent collection highlighted in red.

2 The Case of QREs

In contrast to previous visual REG tasks based on identification of a single object (Mitchell et al., 2010, 2013), QRE tasks involve referring to *collections* of visual items. Two examples of QRE problems are found in Figure 1. Given that the items are homogeneous and randomly arranged,

Proceedings of The 12th International Conference on Natural Language Generation, pages 11–18,
Tokyo, Japan, 28 Oct - 1 Nov, 2019. ©2019 Association for Computational Linguistics

quantity becomes a salient means of describing a target set of items (indicated in red). For example, one could use an *exact number* expression (e.g., "the box with *eighteen* circles" (A) or "the box with *three* circles" (B)). One could also refer to *relative* quantity (e.g., "the box with *more* circles" (A&B)). Other forms of expression include *vague* expressions of quantity (e.g., "the box with *many* circles" (A) or "the box with *a few* circles" (B)) and *absolute* descriptions that refer to the presence of items (e.g., "the box with dots" (A&B)). The reader may find that the sort of expression he or she finds natural differs between problems A and B. It is likely the case that the reader finds it natural to describe the target in (B) with exact number, whereas the reader is more likely to describe the target in (A) in less precise terms. To account for this phenomenon, we turn to the psychophysics of human numerical perception.

The perception of quantity consists of multiple processes, each occurring at different rates and resulting in mental representations of varying precision. Explicit counting provides a slow, but precise, determination of number (Gelman and Gallistel, 1986), in which each visual item requires roughly 250–350 ms to enumerate (Trick and Pylyshyn, 1994). Estimation provides a rapid, but less precise, judgment of the quantity of a group of objects (Barth et al., 2003). A third process exists: subitizing, i.e. the rapid and precise judgment of numerosity for quantities from 1-4 (Kaufman et al., 1949). Within the subitizing range, each visual item requires only 40–100 ms to accurately enumerate (Trick and Pylyshyn, 1994). Thus, accurate *exact* numerical descriptions of small collections of items are fast and easy, whereas such descriptions require time and effort for larger quantities of items, so that *vague*, *relative*, or *absolute* descriptions are quicker and easier to produce for larger quantities.

Human subject experiments in QRE generation (Barr et al., 2013) show that this perceptual effort affects language usage even when there are no time limits in viewing the stimuli or generating QREs. The frequency of QRE types produced by subjects in Barr et al. (2013) study (for participants that gave more than one type of response) is plotted in the top portion of Figure 2. In the problem IDs, the number preceded by a 'T' indicates the target set quantity, while numbers preceded by 'D' indicate distractor set quantities. While the data suggest a general tendency toward exact responses, it is clear that exact responses are significantly preferred in the subitizing range, while dispreferred outside the subitizing range.

One could propose that algorithms used to tackle QREs simply have a rule that treats quantities under four differently. However, studies have shown that common arrangements of visual items, such as the faces of six-sided dice, have been shown to be rapidly and accurately enumerated even beyond the traditional four-item subitizing limit (Mandler and Shebo, 1982) and that subitizing can be disrupted by attentional load (Railo et al., 2008). This suggests that a general solution in adapting current REG algorithms lies not in simple ad hoc rules, but rather in using more comprehensive models of human perception to inform what can or cannot be plausibly perceived given situational constraints, such as exogenous time limitations, or the desire to minimize perceptual effort or cost (self-imposed time limitations).

3 Approach

To apply perceptual cost pruning to QREs, our approach consists of two steps: (1) determine what quantified information can be perceived given a time limit t, and remove information that does not meet this threshold, and (2) apply the Incremental Algorithm to this reduced set of information, ignoring attributes that have been removed.

3.1 Time Cost of Exact Enumeration

Given that estimation occurs rapidly, we focus on modeling the time course of exact enumeration. The literature on numerical perception suggests that the time it takes to exactly enumerate n items follows a bilinear time function (Trick and Pylyshyn, 1994) that can be expressed as follows:

$$T_{exact}(n) \approx T_f \cdot min(r_s, n)$$

$$+ \prod_{max(r_s,n) \leq i \leq n} T_{subvocal}(i) + T_f$$

where T_f denotes the time necessary to attend to encode a single item into visual memory, r_s denotes the subitizing limit, and $T_{subvocal}(i)$ denotes the time necessary to subvocalize the i-th count word. Briggs et al. (2017) present a computational implementation of subitizing and counting that exhibits the above function for exact enumeration response time. In this paper, we draw from Briggs

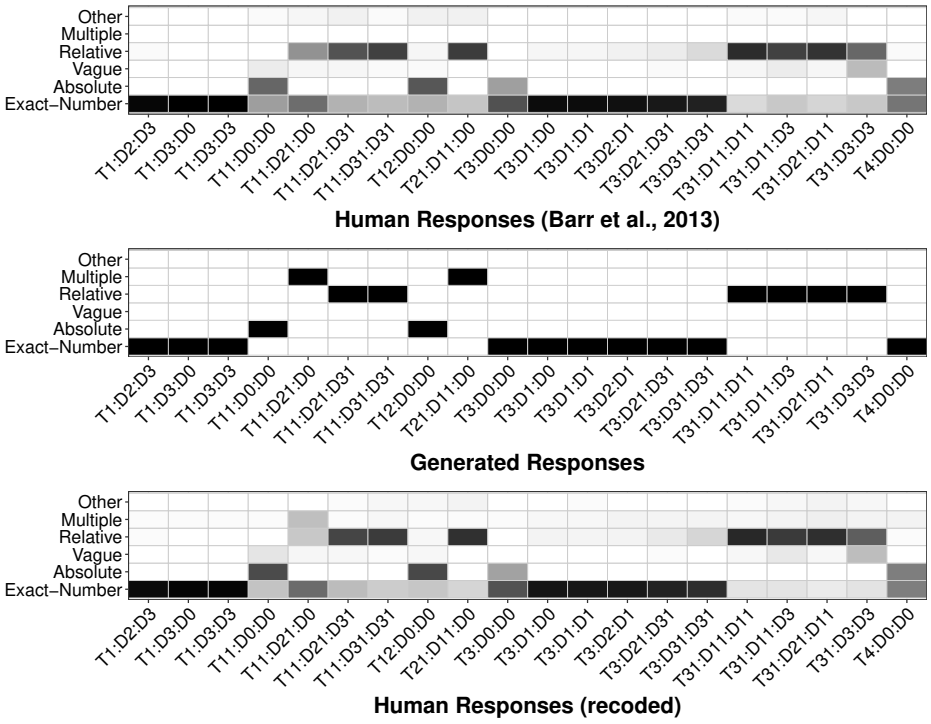

Figure 2: Graphical plot of frequency (darker shading indicates higher frequency) of each category of reference expression (y-axis) for each QRE problem (x-axis) for both human participants from (Barr et al., 2013) (above); the results of the best preference ordering for the Incremental Algorithm (IA_{NAR}) after perceptual cost adjustment of the knowledge base (middle); and recoded responses (bottom). In the problem names, the number preceded by a 'T' indicates the target set quantity, while numbers preceded by 'D' indicate distractor set quantities.

et al. (2017) to set the following values: T_f = 50ms, $r_s = 4$, $T_{subvocal}(i) = 250$ ms, which is shown to be a good fit with human response time data.

3.2 Factoring In Perceptual Cost

The first step in QRE generation is producing a knowledge base for a particular QREG trial. In this paper, we consider three types of quantified expression: absolute (A), exact number (N), and relative (R). The presented classification scheme is different from Barr et al. (2013)'s original scheme in that it consolidates the categories superlative and comparative into the category relative. We leave the production of *vague* quantified expression for future work. This is due in part to the relative infrequency of vague expressions in the human data.[1]

To demonstrate the application of perceptual cost pruning, we consider an example from Barr et al. (2013), whose study was made up of 20 QREG tasks, each consisting of a target (T) and two distractors (D). Each QREG task is specified by the number of items in the target collection and in each distractor collection. Therefore, the exact number information is available without further processing. Absolute attributes are derived by a simple check as to whether or not the number of items in each collection is greater than or equal to zero. Relative attributes are derived by checking whether or not the target or distractor is the collection with the maximum or minimum number of items. If not, then the collection can be labeled as not having the most or least items.

To illustrate, let us consider the problem `T21:D11:D0`, in which the target collection has 21 items and the two distractors have 11 and 0 items, respectively. We can translate the information about each item into the following collection of quantity attributes:

- $Target$ = [A : "has items"; N : "has 21 items"; R : "has the most items"]

[1] Additionally, vague quantified expressions (such as "several" or "many") lack intensional definitions that would enable us to code simple rules to appropriately populate our knowledge base with the correct vague expressions. Additional human data is needed to inform an extensional definition of vague quantified expressions.

- $Distractor1$ = [A : "has items"; N : "has 11 items"; R : "does not have the most items"]
- $Distractor2$ = [A : "has no items"; N : "has 0 items"; R : "has the fewest items"]

Without perceptual cost pruning, the above knowledge base is passed to the REG algorithm as is. However, with perceptual cost pruning, we replace values that cannot be perceived under time constraints with null tokens. For example, if we assume only 0.5s are available to evaluate each collection (either through external time limitations or through deliberate choice to limit gaze time), our model of exact enumeration (above) would indicate that 21 and 11 items cannot be exactly enumerated and therefore would be pruned. This would result in the following modified knowledge base:

- $Target$ = [A : "has items"; N : $\emptyset$; R : "has the most items"]
- $Distractor1$ = [A : "has items"; N : $\emptyset$; R : "does not have the most items"]
- $Distractor2$ = [A : "has no items"; N : "has 0 items"; R : "has the fewest items"]

To handle a pruned knowledge base, we implemented a modified version of the IA, which we describe in the following section.

3.3 Modified IA

Given a target referent T, a set of distractors $D = D_1, ..., D_n$, and a preference ordering of attributes $A = [a_1, ..., a_m]$, the IA selects a subset of attributes to include in a referring expression (RE) by traversing the set of attributes in order of preference, adding the current attribute to the RE if it eliminates any of the remaining distractors (Krahmer and Van Deemter, 2012). Perceptual cost pruning can be accommodated by allowing the algorithm to skip an attribute if the target's value for the current attribute is equal to the null token.

Note that we also assume that if a distractor object's attribute value has been replaced by the null token, but the target's matching attribute value has not been replaced, then the speaker still represents this as a difference that allows for the elimination of that distractor. This can be justified under the simplifying assumption built into our knowledge base pruning method that people would devote equal time to perceive all collections. Therefore, if one collection's quantity was able to be perceived exactly under t seconds, while the other was not, it is evidence against the two collections having the same quantity. However, as this is a preliminary idea, further investigation is needed to establish the limitations of this assumption.

4 Initial Evaluation

To evaluate the effectiveness of the IA with and without perceptual cost pruning, we rely on the Dice coefficient, commonly used in REG tasks, and defined below (Van Deemter, 2016):

$$Dice(H, A) = \frac{2 \times |H \cap A|}{|H| + |A|}$$

where H is the set of attributes found in a human-generated description and A is the set of attributes found in an algorithmically generated description. As previously mentioned, we consider here three types of quantified expression: absolute (A), exact number (N), and relative (R). Combining these three options into all possible orderings generates six preference orderings for the Incremental Algorithm. For example, an IA ordering of $A \succ N \succ R$ is denoted IA_{ANR}.

We calculated Dice scores for the IA for all possible orderings of these expression types. To test perceptual cost pruning, Dice scores were calculated for all six preference orderings for time limits between 0.5s ms and 10s (at 0.1s intervals). As originally reported by Barr et al. (2013), roughly 20% of participants only gave exact number responses. This could reflect different perceptual and generation strategies, specifically the difference between a *fixed* strategy, in which the speaker always takes time to count and reports an exact numerical description, and a *flexible* strategy, in which the speaker takes a more limited time to perceive each item and reports a description based on what was plausibly perceived in the limited timeframe. We predict that perceptual cost pruning will help account for participants that exhibit a flexible strategy. In our analysis, we calculate three Dice scores corresponding to the 20% of participants that exhibited the fixed strategy, the remaining participants that exhibited a flexible strategy, and the combined set of all participants. We report the results from the time ranges found to yield the highest Dice score for perceptual cost pruning in Table 1.

REG Algorithm	Subject Group	No PCP	PCP
IA_{ARN}	Flexible + Fixed (All)	0.300	0.300 ($2.2s \leq t \leq 4.4s$)
	Flexible Only	0.368	0.368 ($0.5s \leq t \leq 1.9s$)
	Fixed Only	0.0	0.0 ($t > 7.0s$)
IA_{ANR}	Flexible + Fixed (All)	0.571	0.635 ($2.2s \leq t \leq 4.4s$)
	Flexible Only	0.535	0.703 ($0.5s \leq t \leq 1.9s$)
	Fixed Only	0.730	0.734 ($t > 7.0s$)
IA_{RAN}	Flexible + Fixed (All)	0.235	0.235 ($2.2s \leq t \leq 4.4s$)
	Flexible Only	0.289	0.289 ($0.5s \leq t \leq 1.9s$)
	Fixed Only	0.0	0.0 ($t > 7.0s$)
IA_{RNA}	Flexible + Fixed (All)	0.235	0.235 ($2.2s \leq t \leq 4.4s$)
	Flexible Only	0.289	0.289 ($0.5s \leq t \leq 1.9s$)
	Fixed Only	0.0	0.0 ($t > 7.0s$)
IA_{NAR}	Flexible + Fixed (All)	0.649	0.706 ($2.2s \leq t \leq 4.4s$)
	Flexible Only	0.568	**0.752** ($0.5s \leq t \leq 1.9s$)
	Fixed Only	**1.0**	**1.0** ($t > 7.0s$)
IA_{NRA}	Flexible + Fixed (All)	0.649	**0.716** ($2.2s \leq t \leq 4.4s$)
	Flexible Only	0.568	0.713 ($0.5s \leq t \leq 1.9s$)
	Fixed Only	**1.0**	**1.0** ($t > 7.0s$)

Table 1: DICE scores for all IA preference orderings with no knowledge base adjustment (left) and with adjustment by perceptual cost pruning (right).

4.1 Naive IA

Because referring to the exact quantity of the target set was sufficient to eliminate distractors, instances of IA_{N**} *without* perceptual cost pruning simply produced all exact number responses. Given that exact number responses were found to be most common in the human data, these orderings were found to have the highest Dice scores. Including all participants, the Dice score of IA_{N**} was 0.649. This score decreased for the subset of participants that produced more than one type of response (Dice score of 0.568), showing that these participants expressed quantity non-exactly in some problems. In particular, these participants used non-exact quantified expressions a majority of the time in 9 out of 20 problems (see Figure 2).

4.2 IA with Perceptual Cost Pruning

Including all participants, the best Dice score (0.716) was found to be associated with a time limit of 2.2-4.4 seconds and the IA_{NRA} preference ordering. With the subset of flexible response participants, the IA_{NAR} preference ordering produced the best Dice score (0.752), during time limits of 0.5-2.0 seconds. In both cases, perceptual cost pruning increased the human-likeness of the IA relative to the baseline IA. As expected, perceptual cost pruning did not improve the similarity scores for subjects with fixed response strategies. More notable, however, is the fact that the IA with perceptual cost pruning would predict that at least 7.0 seconds are needed to view some target collections to obtain an exact numerical description and match the performance of the naive IA. The generated response types for these best runs IA_{NAR} with perceptual cost pruning are plotted in the middle chart of Figure 2. In these runs, the majority human response was correctly predicted 18 out of 20 times.

It is worth noting those instances where the predictions diverge from the human data. IA_{NAR} predicts an RE with both absolute and relative descriptors, (i.e., 'Multiple') on problems such as `T21:D11:D0` and `T11:D21:D0`. Because a brief perceptual time limit of less than two seconds is insufficient to exactly enumerate 21 or 11 dots, IA_{NAR} would predict that no exact descriptors would be used. However, because one of the distractors is empty and the other is non-empty, an absolute descriptor is added to the RE, but is still insufficient to eliminate all the distractors.

Our model's prediction for 'Multiple' expression types differs from Barr et al. (2013), as they did not originally annotate responses as having more than one type of quantified expression (so

they could not have any aggregate 'Multiple' category for expressions with multiple types of descriptors). This additional category provides an opportunity to plot when our modified IA algorithm with perceptual cost pruning would predict more than one type of quantified descriptor. For example, some of the human-generated responses were consistent with a 'Multiple' coding, e.g., in problem `T11:D21:D0` with expressions such as "the square with some triangles but not the most triangles," and "the square containing the smaller number of symbols but not the blank square."[2] However, the majority of the human-generated responses for a task like `T21:D11:D0` were *not* consistent with a 'Multiple' coding; the majority response was simply a relative expression. Since our algorithm generated descriptions according to a scheme that did not match Barr et al. (2013)'s original coding, we re-examined and recoded a subset of their original data.

5 Recoding and Re-evaluation

After surveying the annotations from Barr et al. (2013), we decided to recode their data for two key reasons. First, as described above, the original coding marked responses as having only one expression type, whereas various expressions in the data were found to be plausibly ascribed two or more descriptor types. Second, the original coding was performed automatically by pattern-matching and had incorrectly labeled a significant portion of responses as exact number expressions (e.g., "the one with the most dots"). The patterns we identified for potential error and recoding are as follows: "one with", "other two', and "out of the three". In such cases, the numbers "one", "two", and "three" refer to the squares containing the collections of visual objects and not the quantity of items they contain. In addition, we verified the coding for all responses first marked OTH ("Other").

In total, we targeted for recoding 491 of the 1508 responses Barr et al. (2013) reported. Our recoding was performed by two annotators, who began by identifying expressions that contained multiple types of expressions. Agreement was calculated using Kendall's coefficient of concordance, which indicated high interannotator agreement ($W = .906$) regarding which expressions contained more than one type of expression. Automated pattern matching was used to identify expressions that contained the template strings specified above. High interannotator agreement scores (Kendell's coefficient of concordance $W > .95$) were found across all expression types (ABS, NUM, BASE, and OTH).

[2]Data publicly available at: https://staff.fnwi.uva.nl/r.fernandezrovira/xprag/

REG Algorithm	No PCP	PCP
IA_{ARN}	0.411	0.411
IA_{ANR}	0.536	0.732
IA_{RAN}	0.312	0.312
IA_{RNA}	0.312	0.312
IA_{NAR}	0.530	**0.773**
IA_{NRA}	0.530	0.717

Table 2: DICE scores under recoded data for all IA preference orderings with no perceptual cost pruning of the knowledge base (left) and with knowledge base adjustment by perceptual cost pruning (right).

5.1 Recoding Results

A plot of the recoded human responses can be found at the bottom of Figure 2. One major result the revised coding yielded was a revision of the number of participants that produced only one QRE type throughout the task. Barr et al. (2013) reported that 20% of participants only generated QREs with exact number descriptions. However, after recoding, only three participants (6% of all participants) were shown to exclusively use exact number descriptions, i.e., to rely on a fixed REG strategy.

Our recoding supported our initial observations regarding problem `T11:D21:D0`. Although exact number REs are still the plurality response for this problem, the plot in Figure 2 does indicate that the complex RE with multiple quantified expression types is the second most common response. As predicted by the IA_{NAR} preference ordering, nearly all of these complex REs use both absolute and relative descriptions. One possible reason that an exact number response is still frequent for this problem, but not for `T21:D11:D0`, is that the 11 visual items representing the target is at the threshold of what people can quickly judge to be "countable" (Mandler and Shebo, 1982). In contrast, 21 items are judged to not be easily "countable."

5.2 Naive IA

Because the revised number of respondents that used only one type of RE was significantly re-

duced, we do not report Dice scores for the different subsets of participants that used flexible or fixed REG strategies. Instead, we simply report the aggregate Dice score for all participants. As many previous instances of exact number responses were reexamined and revised, the Dice scores of IA_{N**} preference orderings are reduced from 0.649 to 0.530. Conversely, the smaller number of exact number expressions in the revised annotations improves the performance of the IA_{A**} and IA_{R**} preference orderings: their Dice scores increase with the new coding scheme. In contrast to the original coding, the best preference ordering without perceptual cost pruning is IA_{ANR}, with a Dice score of 0.536.

5.3 IA with Perceptual Cost Pruning

With perceptual cost pruning, the best Dice score (0.773) was found to be associated with a time limit of 0.2-1.9 seconds and the IA_{NAR} preference ordering. This reflects a better fit of the IA with perceptual cost pruning to the human data under the revised coding (compared with the best Dice score of 0.706 for the original coding). Additionally, the Dice score of the naive IA also improved under the new coding scheme. Under the original coding, the aggregate Dice score for all participants increased by only 0.057 for IA_{NAR} and 0.067 for IA_{NRA}. In contrast, the revised codings yield a Dice score improvement of 0.243 and 0.187 for IA_{NAR} and IA_{NRA}, respectively.

6 Discussion and Future Work

We have shown that the ability of a preference order based REG algorithm (IA) to produce human-like responses in QRE generation tasks can be significantly improved by integration of a model of perceptual cost. We believe this success can be attributed to the fact that our perceptual cost pruning approach captures the underlying tension in QRE generation between the desire to be as informative as possible and the desire to minimize perceptual effort, which reduces the precision of information (Barr et al., 2013). The desire to be informative is evidenced by the primacy of exact numbers in the best fit preference orderings, whereas the desire to minimize perceptual effort is evidenced by the fact that the time limit of the best fitting IAs is $\leq$2 seconds. The tradeoff between the desire to be as informative as possible and perceptual limitations has also been shown in recent work, where the precision of quantified descriptions of visual scenes decreases as presentation time decreases (Briggs et al., 2019).

Furthermore, our presented approach leads to a variety of predictions that we are testing. One such prediction involves the role of a set of items' spatial arrangement on QRE production. Findings in numerical perception indicate that common arrangements of visual items, often deemed canonical patterns, can be exactly enumerated more quickly than randomized patterns (Mandler and Shebo, 1982; Wender and Rothkegel, 2000). This would suggest that a greater amount of exact number REs would be produced for well-learned patterns (such as dice faces), since their enumeration is perceptually cheap. On the other hand, it is possible that canonical patterns would be described as canonical, without any reference to number (e.g., "the square with the dice pattern"). This raises the larger question of what attributes and descriptions of a group of visual items are preferred over one another.

Additionally, we wish to investigate how the psychophysics of approximate number representations may also limit what quantity information is available for REG. In the present study, large quantities were sufficiently different so that estimation easily yields valid relative comparisons. However, some numerosity differences are not so easy to assess by estimation (e.g., imagine attempting to tell the difference between 62 and 64 objects without counting). A complete account of QREG must include both the limitations of exact and inexact numerical representation.

Finally, if additional experiments are designed such that the quantity of the total number of individuals in a collection does not provide a clear differentiating attribute, we would predict that other properties of visual groups would be referenced. Specifically, our predictions are that the following properties could be used:

Spatial descriptions of the group - groups of visual items can be described across a variety of spatial dimensions that are unrelated to the quantity of items in the group. Examples of types of spatial descriptions could include: area, shape, and density of the cluster of visual items.

Number of subgroups - people have the ability to group visual items together based on proximity (Im et al., 2016). In other words, people can not only refer to the total number of individuals in a

collection, but also to the number of subgroups within a collection or the size of these subgroups.

Further data collection is needed to determine which forms of these visual group properties are commonly generated and which are preferred over one another.

7 Conclusion

In this paper, we investigated the problem of generating referring expressions in visual contexts based on differences in the quantity of a target collection and distractor collections. Given the low performance of a traditional, preference ordering based REG algorithm in this task, we have demonstrated the importance of factoring in perceptual cost in REG. We have also proposed and validated a novel method, called perceptual cost pruning, of factoring in perceptual cost by ablating a knowledge base according to models of human psychophysical limits. Future work is needed to further refine this proposed method and explore REG in the context of differentiating collections of visual items with varying spatial arrangements.

Acknowledgments

This work was supported by an NRL Karles Fellowship awarded to the first author, an NRC Postdoctoral Fellowship awarded to the second author, and AFOSR MIPR grant F4FGA07074G001. The views expressed in this paper are solely those of the authors and should not be taken to reflect any official policy or position of the United States Government or the Department of Defense.

References

Dale Barr, Kees van Deemter, and Raquel Fernández. 2013. Generation of quantified referring expressions: evidence from experimental data. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 157–161.

Hilary Barth, Nancy Kanwisher, and Elizabeth Spelke. 2003. The construction of large number representations in adults. *Cognition*, 86:201–221.

Gordon Briggs, Will Bridewell, and Paul F. Bello. 2017. A computational model of the role of attention in subitizing and enumeration. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, pages 1672–1677, London, UK.

Gordon Briggs, Christina Wasylyshyn, and Paul F. Bello. 2019. Elicitation of Quantified Description Under Time Constraints. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, pages 1436–1442, Montreal, Canada.

Alasdair Clarke, Daniel Francis, Micha Elsner, and Hannah Rohde. 2013. Where's wally: the influence of visual salience on referring expression generation. *Frontiers in Psychology*, 4:329.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19:233–263.

Micha Elsner, Alasdair Clarke, and Hannah Rohde. 2018. Visual complexity and its effects on referring expression generation. *Cognitive Science*, 42:940–973.

Rochel Gelman and Charles R Gallistel. 1986. *The child's understanding of number*. Harvard University Press, Cambridge, MA.

Hee Yeon Im, Sheng-hua Zhong, and Justin Halberda. 2016. Grouping by proximity and the visual impression of approximate number in random dot arrays. *Vision Research*, 126:291–307.

Edna L Kaufman, Miles W Lord, Thomas Whelan Reese, and John Volkmann. 1949. The discrimination of visual number. *The American Journal of Psychology*, 62:498–525.

Emiel Krahmer and Kees Van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38:173–218.

George Mandler and Billie J Shebo. 1982. Subitizing: An analysis of its component processes. *Journal of Experimental Psychology: General*, 111:1–22.

Margaret Mitchell, Kees van Deemter, and Ehud Reiter. 2010. Natural reference to objects in a visual domain. In *Proceedings of the 6th International Natural Language Generation Conference*, pages 95–104. Association for Computational Linguistics.

Margaret Mitchell, Kees Van Deemter, and Ehud Reiter. 2013. Generating expressions that refer to visible objects. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics (ACL).

Henry Railo, Mika Koivisto, Antti Revonsuo, and Minna M Hannula. 2008. The role of attention in subitizing. *Cognition*, 107:82–104.

Lana M Trick and Zenon W Pylyshyn. 1994. Why are small and large numbers enumerated differently? a limited-capacity preattentive stage in vision. *Psychological Review*, 101:80–102.

Kees Van Deemter. 2016. *Computational Models of Referring: A Study in Cognitive Science*. MIT Press, Cambridge, MA.

Karl F Wender and Rainer Rothkegel. 2000. Subitizing and its subprocesses. *Psychological Research*, 64:81–92.