Visual grouping and pragmatic constraints in the generation of quantified descriptions

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

Abstract

Studies suggest that people use the least possible effort to generate natural language descriptions of sets of objects. This means that they base descriptions on what is perceptually available to them. For instance, people can subitize, i.e., rapidly assess the exact quantity of small numbers of objects, so when the quantity of objects in the visual scene is beneath this threshold, they give numeric descriptions; when the quantity is above this threshold, they generate non-numeric descriptions. However, no research examines how people describe visual scenes of items in groups. As such, it is unclear how people will form descriptions of scenes that contain a large total number of items in groups. We report on a novel experiment designed to investigate how people produce quantified descriptions of scenes composed of salient visual groups. The results corroborate the least effort hypothesis, and suggest that people's incremental perception of quantity drives their descriptions.

Keywords: numerical perception; pragmatics; quantified description; subitizing; visual grouping

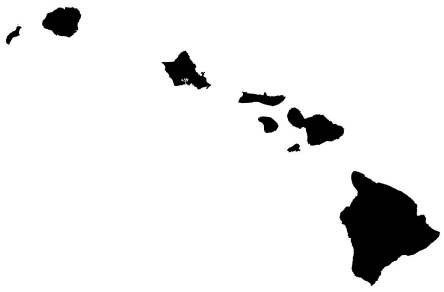

Figure 1: Outlines of major islands comprising the U.S. state of Hawaii: an example of a collection of distinct visual objects.

Introduction

People make use of quantified descriptions to characterize visual scenes. Consider a map of the state of Hawaii (see Figure 1) and imagine how you might describe the number of islands that compose it. You could describe it using exact numbers (e.g. "I see *exactly eight* islands"); bounded numerals (e.g., "I see *more than five* islands"); or vague quantifiers (e.g. "I see *several* islands."). Pragmatic and discourse goals will likely affect the level of precision you use to describe the total number of islands (Cummins, 2015; Hesse & Benz, 2018), and limited attentional and perceptual resources likewise affect the precision with which people describe quantities (Briggs, Wasylyshyn, & Bello, 2019).

People rapidly and accurately determine the number of small groups of items in a process called *subitizing* (Kaufman, Lord, Reese, & Volkmann, 1949). While there might be some variation to the subitizing limit, it is safe to say that people can subitize quantities up to 4 (Mandler & Shebo, 1982). Exact enumeration within the subitizing range requires between 40–100 ms for each visual item (Trick & Pylyshyn, 1994). For sets of items that fall outside the subitizing range, people appear to have two ways of assessing quantity. First, they can estimate quantity using a mental representation known as the approximate number sense (ANS). The ANS refers to a developmentally primitive representational capacity that allows people to perform numerical calculations without assigning names to numbers. Hence, the ANS does not provide access to individual items or yield exact numbers, but people can use it to make comparisons, such as estimating whether one group of items is bigger than another (Dehaene, 1992; Izard & Dehaene, 2008). Second, people can process sets of items out of the subitizing range by counting them, which allows them to construct more precise quantity representations. Counting demands cognitive effort and can be slow: people count items at a rate of about 250–350 ms per item (Trick & Pylyshyn, 1994).

Empirical investigations into how people perceive numerosity tend to adopt restrictive tasks that do not allow people to generate a broad range of descriptions. For instance, many numerical perception tasks force participants to report the exact number of items in a scene (e.g., Mandler & Shebo, 1982). Initial work in more free-form quantified description tasks required participants to describe a set of items in contrast to an alternative set. In daily life, participants do not abide by such constraints: your description of the Hawaiian archipelago did not depend on comparisons to other archipelagos, and a description, such as, "a bunch of islands" may suffice. Few studies examine how people naturally formulate quantified descriptions of visual scenes.

Recent work suggests that when placed under time constraints, people describe sets of items by appealing to inexact language (Briggs et al., 2019). Barr, van Deemter, and Fernández (2013) also showed that people use exact numbers to describe items below the subitizing range and

non-numerical quantifiers for items above it. They examined how people use quantified expressions to distinguish one set of objects from other sets of different quantities. Participants used numbers 90% of the time to distinguish sets with a subitizable number of objects, but did so only 39% of the time for the targets with numerosity beyond the subitizing range. They often used numeric descriptions when comparing against sets where the distractors were blank images: the result would seem to violate the Gricean maxim of quantity (Grice, 1975), since reporting the number of items is overinformative when a simpler response, e.g., "the one with the circles" would suffice. Barr et al. (2013) hypothesized that people balance informativity with the effort required to produce a description. Briggs and Harner (2019) interpreted the proposal to concern the information produced by perceptual processes in limited time frames. We synthesize both proposals as follows:

> **Least effort hypothesis**: speakers generate descriptions using as little effort as their processing limitations permit. Their descriptions may be overinformative in situations in which perceptual processes make salient information – such as numerosity, color, and shape – accessible.

The hypothesis explains participants' preference to produce descriptions such as "a set of four circles" when "a set of circles" would suffice.

People need not rely on descriptions of total quantity to characterize groups of items at all. For example, participants in Barr et al. (2013) also produced referring expressions that used spatial information such as density to discriminate targets: one participant generated an expression that described a target as "the most crowded." When describing the Hawaiian islands (Figure 1), you might describe its overall shape (e.g., "a group of islands lying along a curved path"). Likewise, you could break the island chain into subgroups and describe the number of subgroups ("three groups of islands") or the cardinality of each subgroup ("a group of two islands, a group of five islands, and a single large island"). People can refer to the number of subgroups and the cardinality of each subgroup, either by themselves (e.g., "three groups") or as an addendum to their description of the total quantity of items (e.g., "eight islands in three groups").

Barr et al. (2013) and Briggs et al. (2019) investigated how people produce quantified descriptions of collections of objects in scenes without salient groups. In this paper, we report investigations into how people's quantified descriptions change in the presence of salient groups. We begin by reviewing prior work on the effects of visual grouping on enumeration and discuss how visual grouping may interact with pragmatic constraints to affect quantified descriptions. We then describe an experiment designed to elicit quantified descriptions of visual scenes under different grouping conditions. Finally, we discuss how our results relate to the tension between perception and pragmatics in visual description.

Perceiving and describing groups

Grouping and the spatial arrangement of objects affect the perception and representation of numerosity. People perceive grouped items as more numerous than those that are not grouped (Poom, Lindskog, Winman, & Van den Berg, 2019; Vos, Van Oeffelen, Tibosch, & Allik, 1988). Cantrell and colleagues (2013; 2015) studied the effects of grouping on the perception of individuals; they found that people consider the features of groups more prominent than the features of the individual group members. Items in more regular patterns likewise appear more numerous than randomly positioned items (e.g., Ginsburg, 1976, 1980; Cousins & Ginsburg, 1983). Similarly, items organized into groups via Gestalt principles appear more numerous than groups without such organization (Frith & Frit, 1972). The amount of space that items occupy affects perception of their numerosity: elements that take up more space appear more numerous than those that take up less (Vos et al., 1988)

Research also demonstrates that visual grouping makes enumeration easier. For instance, people have an easier time estimating the numerosity of a set of dots distributed into regular patterns than the same set of dots distributed randomly (Burgess & Barlow, 1983). Van Oeffelen and Vos (1982) found that people are faster at giving overall item numbers outside the subitizing range if they can break the scene up into a subitizable number of groups. The findings suggest that people can quickly perceive subitizitable groups, evaluate the individual cardinalities of their groups, and sum those cardinalities to estimate the total number of items in a scene (Fernberger, 1921; Oberly, 1924). Starkey and McCandliss (2014) refers to the process as "groupitizing."

Research into how people form quantified descriptions of scenes ignores how groups affect descriptions. Above, we identified three relevant referents when considering quantified descriptions of groups: the total number of visual items, the number of groups, and the cardinality of each group. Consider the following possible quantified descriptions of a hypothetical scene that depicts 3 groups of 4 dots:

"There are 12 dots." [**D1**]

"There are 12 dots in 3 groups." [**D2**]

"There are 3 groups of 4 dots." [**D3**]

"There are 12 dots in groups of 4." [**D4**]

"There are 3 groups of 4 dots for a total of 12 dots." [**D5**]

D1 concerns the total quantity of items in the scene; D2 concerns the total items and the number of groups; D3 presents the number of groups and the total number of items; D4 presents the number of groups and their cardinality. And D5 presents all three sorts of information. D1 is underinformative: it doesn't provide any information that the dots are organized into groups. D2 is also underinformative, since it is compatible with 12 items separately into groups of non-uniform cardinality. D3 and D4 are equally informative. D5

is overinformative: of the three numbers it provides, one of the numbers can be computed from the other two.

As we will show, over 95% of people's responses fall into one of the five patterns above. Which description do theories predict people generate to describe quantities of items? The Gricean maxim of quantity suggests that people should not be underinformative – e.g., they shouldn't describe 3 groups of 4 dots with descriptions such as, "there are 3 groups," since the description is compatible with, say, a set of 3 groups of 100 dots. Likewise, they shouldn't produce descriptions such as D1 and D2, which are similarly underinformative, or D5, which presents redundant information. But the maxim of quantity provides no guide on how to appropriately describe groups and their composition: to describe the scenario above, is it more sensible to say, "there are 3 groups of 4 dots" (D3) or "there are 12 dots in groups of 4" (D4)?

It is not surprising that pragmatic accounts make no distinction between D3 and D4, as they do not depend on processing constraints. But the two descriptions are equally informative, even though they differ in complexity – to infer the missing information, D3 requires multiplication, and D4 requires division. The least effort hypothesis, introduced earlier, concerns people's processing constraints in perceiving visual scenes. In applying it to images with visually salient groups, it predicts that people's descriptions will depend on whether they concern subitizable groups of items. When a set of subitizable groups is present, speakers should have little difficulty perceiving both the number of groups and their cardinality. But, unless the total number of items can be subitized, it should be inaccessible to individuals unless they carry out mental arithmetic. The least effort hypothesis therefore predicts the following:

> **Least effort hypothesis predictions:** For scenes that depict a subitizable set of items, speakers should generate descriptions such as D1. In contrast, for scenes that depict a non-subitizable set of items arranged into groups, speakers should generate descriptions such as D3 more often than any other description.

Barr et al. (2013) did not explain the mental computations that underlie the least effort hypothesis. Here, we propose the idea is compatible with the incremental processing of visual scenes. The least effort hypothesis would suggest that, for scenes that can be groupitized, people perceive and encode the number of groups and the cardinality of each group before they determine the total quantity of items in the scene. In fact, people often decompose a cluttered scene into subgroups, which can facilitate mental arithmetic (Starkey & McCandliss, 2014; Ciccione & Dehaene, 2020).

To test the least effort hypothesis, we ran a study that required participants to generate a description from a single image of a collection of objects. Namely, it sought to test whether people's descriptions differed when they described small sets of items in the subitizing range, i.e., sets of 3 or 4 items, and when they described large groups of items in the non-subitizing range, i.e., sets of 9 or 12 items. It further sought to test whether visually grouping sets of items shifted the sorts of quantified descriptions people produced.

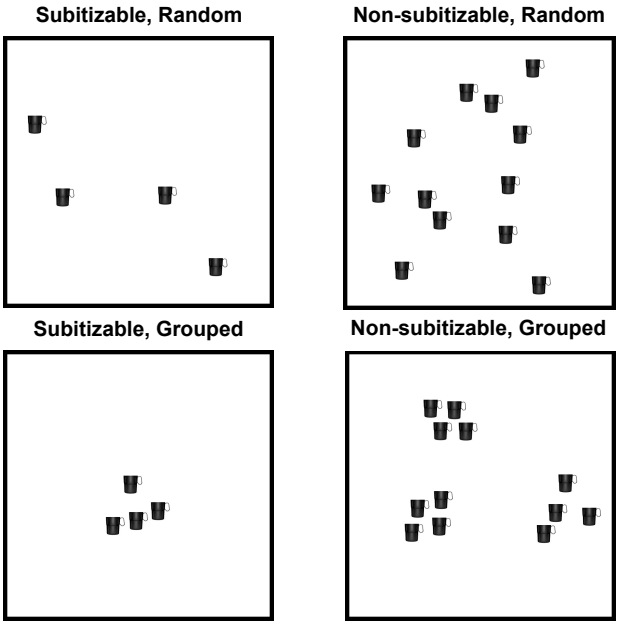

Figure 2: Examples of images from each of the four conditions in the Experiment. Non-subitizable images depict items whose total number cannot be subitized.

Experiment

The experiment tested how people generate natural language descriptions of an image depicting a set of objects. Half of the images presented total quantities of items that people could subitize (i.e., they included either 3 or 4 objects) and the other half presented scenes that had total quantities outside the subitizing range (i.e., they included 9 or 12 objects). And half the images spatially arranged the objects into groups, while the other half scattered the objects in random positions across the image. Participants had to describe the scene as they saw fit.

Method

Participants. Fifty participants (mean age = 38.9 years; 27 males and 23 females) volunteered through the Amazon Mechanical Turk platform (see Paolacci, Chandler, and Ipeirotis (2010), for a review). All participants reported being native English speakers; we dropped five participants for either giving uninterpretable descriptions, or else for giving multiple descriptions on the same trial (e.g., "I see 3 dots. I see more than 2 dots...").

Design. The study manipulated whether images depicting sets of items were grouped or not and whether people could subitize the total number of items or not; the manipulations

yielded a 2 x 2 repeated-measures design. Half the images depicted sets of items that fell within the *subitizing* range, so that images showed 3 or 4 items in total; and the other half showed sets that could not be subitized, i.e., sets that 9 or 12 items in total. And, for half the images showed items in random positions, and the other half showed items in 3 or 4 groups of 3 or 4 objects each.

Procedure. Participants carried out 8 trials, i.e., 2 trials in each of the 4 conditions. The experiment randomized the order of the trials. On each trial, participants viewed a single image and then typed out a description of the scene. Each trial was self-paced with no time limit. Participants typed out their responses to the prompt, "Please enter a description of the image."

Materials. Images in the study concerned sets of everyday objects (e.g., paperclips, mugs, or flowers). Each image depicted one type of object. Figure 2 provides an example of a set of mugs in each of the four conditions in the study.

Data analysis. Three annotators coded participants' responses along categories designed to examine what quantified information they included in their descriptions. Annotators coded each description on whether they mentioned:

1. *The total number of individuals:* whether a response mentioned a numeral to characterize the entire group of objects (e.g., "12 mugs").
2. *The number of groups:* whether a response mentioned a numeral to characterize the number of groups (e.g., "4 groups of mugs").
3. *The cardinality of groups:* whether a response mentioned a numeral to characterize the number of objects in a group (e.g., "groups of 4 mugs").

Participants were free to mention any combination of the quantities above, and so we can classify their responses into $2^3 + 1 = 9$ separate categories, which includes the category of response that lacked any numerals. Despite the diversity of options, participant responses could be classified into one of the five sorts of description introduced above (D1-D5), where D1 concerns any response that mentions the total number of individuals but not the number of groups or the cardinality of groups, and so on for D2-D5. We focus the results on these five relevant patterns of responses.

In addition to coding participants' use of numerals, annotators coded responses on two exploratory measures:

4. *Non-numerical quantifiers:* whether participants used quantifiers that made no use of numerals to refer to either the total number of objects, the number of groups, or the number of objects in each group (e.g., "a few mugs").
5. *Spatial relations:* whether participants made some reference to the spatial arrangement of the collection, e.g. "4 mugs grouped *together*."

Code	Kendall's coefficient of concordance
Total number of individuals	.77
Number of groups	.87
Cardinality of groups	.92
Inexact quantifiers	.88
Spatial relations	.99

Table 1: Three annotators coded participants' responses in the Experiment in five separate ways blind to the condition of the trial. All three coders evaluated the same sample (1/4th) of responses. The table provides their interrater reliabilities. The annotators resolved discrepancies, refined their coding criteria, and then divided the remaining responses to code individually.

Open science. Data from the experiment, experimental code, annotator codings, and statistical analyses are all available online through the Open Science Framework (https://osf.io/u7gyc/).

Results

Figure 3 shows the proportion of responses that fell into the four different categories (D1-D5) as a function of whether images showed items in groups or not and as a function of whether they showed a subitizable set of items or not. Henceforth, we refer to images that depict more items than can be subitized as *non-subitizable images*. We subjected the table of frequencies that underlies Figure 3 to a χ^2 test, which revealed that the frequencies of the four different responses depended on the four conditions in the study ($\chi^2 = 164.03, df = 12, p < .0001$). To perform direct tests of the predictions of the least effort hypothesis, we dummy-coded D1-D5 to treat each one as a binary dependent variable, e.g., 1 if a particular response could be considered D1, and 0 otherwise. Participants made D1 responses far more often (72% of the time) than D2 responses (5%), D3 responses (12.5%), D4 responses (2%); and D5 responses (7%; Wilcoxon tests, zs > 5.67, ps $<$ .0001, Cliff's δs $> .96$). For brevity, the remaining analyses concern the patterns of D1 and D3 responses, but we provide a full battery of statistical analyses online.

As Figure 3 reveals, participants generated D1 responses more often for subitizable than non-subitizable images (86% vs. 56%, respectively; Wilcoxon test, $z = 5.21, p < .0001$, Cliff's $\delta = .63$). And they generated more D1 responses for randomized images than for grouped images (92% vs. 51%, respectively; Wilcoxon test, $z = 5.55, p < .0001$, Cliff's $\delta = .74$). Their tendency to generate D1 responses yielded a reliable interaction: participants tended not to generate D1 responses for non-subitizable images (Wilcoxon test, $z = 4.54, p < .0001$, Cliff's $\delta = .57$). The results corroborate the first of two predictions of the least effort hypothesis: for subitizable images (i.e., the two white bars in the first panel of Figure 3), people provided terse descriptions of the infor-

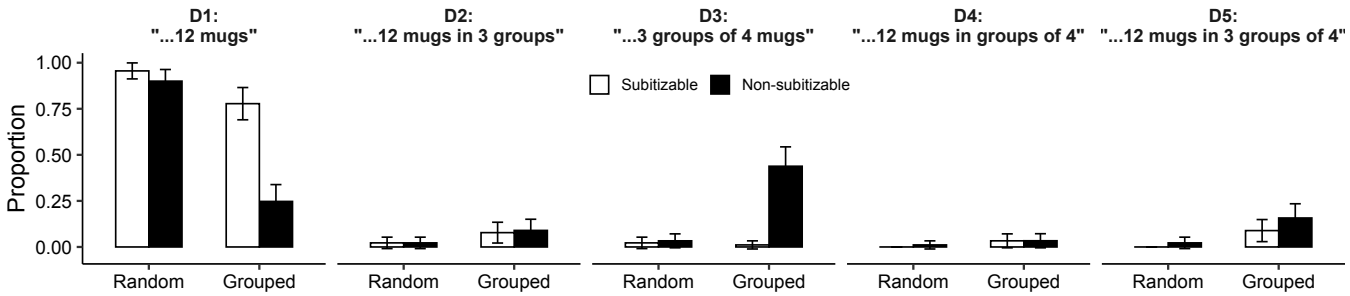

Figure 3: Proportion of participants' descriptions of images in the Experiment that fell into one of five different sorts of description (D1-D5) as a function of whether the image showed a subitizable set of items or not, and as a function of whether salient groups were visible or not. The five sorts of description contained a combination of numeric quantifiers that concerned the total number of items, the number of groups, or the cardinality of groups. Error bars show 95% confidence intervals.

mation for which they have available, i.e., the total number of items. When items are non-subitizable and randomly dispersed, people may have to count them individually to ascertain the total number. But when groups are salient in a non-subitizable image, their strategy shifted.

Participants' pattern of D3 responses revealed this shift in strategy (see the second panel of Figure 3). They made D3 responses 44% of the time for grouped, non-subitizable images. The result yielded a reliable interaction (Wilcoxon test, $z = 4.40, p < .0001$, Cliff's $\delta = .49$). And it yielded reliable main effects, i.e., people produce more D3 responses for subitizable than non-subitizable images (24% vs. 2%, respectively; Wilcoxon test, $z = 4.73, p < .0001$, Cliff's $\delta = .47$); and they produced more D3 responses for grouped vs. randomized images (22% vs. 3%; Wilcoxon test, $z = 4.40, p < .0001$, Cliff's $\delta = .49$).

Planned comparisons show that participants produced more D3 responses than D2, D4, and D5 responses (43% vs. 9%, 3%, and 17%, respectively; Wilcoxon tests, zs > 2.52, ps $< .02$, Cliff's δs $> .31$). Likewise, they produced more D2 responses than D1 responses for grouped, non-subitizable images, though the difference was not statistically reliable (43% vs. 24%; Wilcoxon test, $z = 1.67, p = .09$, Cliff's $\delta = .22$). These results corroborate the remaining prediction of the least effort hypothesis. They suggest that participants based their quantified descriptions, not on considerations of informativity, but rather on perceptual constraints such as the information available to them as they rapidly assessed each image.

Exploratory analyses. In addition to examining participants' usage of numerical quantifiers, we coded their responses along two exploratory dimensions of interest: whether participants used non-numerical quantifiers, e.g., "some," "a few," and "several," and whether they used spatial language, e.g., "grouped close by one another" and "clustered together." In total, participants used non-numerical quantifiers in their responses 7% of the time, and they used spatial language 17% of the time. Figure 4 presents the proportions of responses for which participants used non-numerical quantifiers and spatial language as a function of the conditions in the study.

As the figure shows, participants tended to use non-numerical quantifiers more often for non-subitizable images than for subitizable images (13% vs. 2%; Wilcoxon test, $z = 3.23, p = .001$, Cliff's $\delta = .23$). And they used such quantifiers more often for random images (9%) than for grouped images (6%), but the difference was not reliable (Wilcoxon test, $z = 1.68, p = .09$, Cliff's $\delta = .11$). Nevertheless, the interaction between the two manipulations was significant (Wilcoxon test, $z = 2.35, p = .02$, Cliff's $\delta = .13$).

Participants tended to use spatial language more than twice as often as they used non-numerical quantifiers. Figure 4 shows that for randomized images, their generation of spatial language was nearly identical between subitizable and non-subitizable sets, but for images that depicted salient groups, people used spatial language far more often for subitizable sets than non-subitizable sets; the pattern yielded a significant interaction (Wilcoxon test, $z = 2.98, p = .003$, Cliff's $\delta = .25$), a main effect of subitizability (Wilcoxon test, $z = 2.54, p = .01$, Cliff's $\delta = .18$), and no detectable difference between grouped and randomized images (Wilcoxon test, $z = 0.00, p = .99$, Cliff's $\delta = .02$).

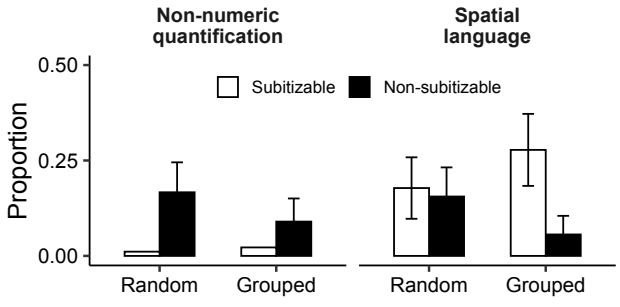

Figure 4: Proportion of participants' usage of non-numerical quantifiers and spatial language in the Experiment as a function of subitizablity and grouping condition. Error bars show 95% confidence intervals.

General Discussion

According to the Gricean maxim of quantity, ideal descriptions of scenes should be neither underinformative – they should not omit relevant and salient information – nor overinformative – they should not be redundant (Grice, 1975). But people flout the maxim systematically (e.g., Okanda, Asada, Moriguchi, & Itakura, 2015), particularly when they generate descriptions of scenes. For instance, they flout the maxim of quantity when they describe an image as "4 black dots" when a description such as "a set of dots" suffices (Barr et al., 2013). To explain such behavior, recent theorists argue that perceptual costs associated with visual properties guide natural language description (Briggs & Harner, 2019; Van Deemter, 2016; Krahmer & Van Deemter, 2012). People generate descriptions in a way that minimizes such costs, a proposal we refer to as the *least effort hypothesis*. The least effort hypothesis argues that speakers generate descriptions based on the information readily available to them. It suggests that people should describe images depicting groups in systematic ways: they should tend to base descriptions on perceptually salient features instead of actively negotiating a balance between informativity and relevance. Hence, it explains why people generate overinformative descriptions. We extended the proposals to concern images depicting groups: because people appear to enumerate small numbers of groups with ease, they should base descriptions on, e.g., the number of those groups instead of the total number of items.

To test the idea, we conducted an experiment that asked participants to generate descriptions of grouped and ungrouped sets of objects. Participants were free to describe sets of items in any way they chose. They responded systematically: most of the time, they used numerals to describe the total number of items on the screen. They did so for subitizable images, i.e., those that depicted sets whose quantities they could rapidly establish, and they did so for sets of higher quantity that were distributed randomly across the image, presumably because they counted the items. But their strategy shifted qualitatively for images that arranged many items into salient groups: participants based their descriptions of such images on the number of groups and the cardinality of those groups, i.e., numbers hypothesized to be accessible to participants without relying on a counting strategy. The results corroborated the predictions of the least effort hypothesis.

The least effort hypothesis may be the end result of the constraints and processes that underlie numerical perception. Such processes are incremental and capacity-limited, i.e., perceivers must devote attentional resources to comprehend complex scenes that depict multiple objects. To alleviate the burden of processing objects serially, people can decompose scenes into subgroups, whose individual members they can then enumerate using arithmetic (Starkey & McCandliss, 2014; Ciccione & Dehaene, 2020). The results we report are consistent with such behavior, and they suggest that the least effort hypothesis is harmonious with incremental perceptual processes: people appear to base their descriptions of images on the earliest available information sufficient to characterize the scene.

One limitation of the present task is the pragmatic demands placed on the speaker. Participants in the study had to generate descriptions by typing them out, and many may have opted for shorter descriptions. The medium may explain why they preferred descriptions such as, "12 items" to descriptions such as "12 items in 3 groups of four," but it does not explain the shift in descriptive strategy for images depicting groups of non-subitizable items. Nevertheless, cooperative speakers should aim to be both informative and brief. One way of construing the least effort hypothesis is as a heuristic by which people achieve that balance without engaging in deliberative counting or arithmetic processes.

One avenue for future work would be to examine scenes with subgroups of heterogenous cardinality. For instance, consider an example of a collection of items with four subgroups each with a different number of items. In this case, it is impossible to concisely provide a description of the cardinality of each subgroup; the least effort hypothesis predicts that people should not describe group cardinality. Another test of the hypothesis could place speakers under time constraints, both for viewing the stimuli and for generating descriptions. Participants would generate more terse descriptions, but the information they choose to omit may prove informative. For instance, limited presentation time may make it difficult to assess the cardinality of groups and drive participants to base descriptions on the number of groups alone.

To build interpretable image description systems, researchers must first understand how people perceive scenes. Many contemporary natural language generation algorithms assume a fully explicit, symbolic representation of a visual scene (e.g., Krahmer & Van Deemter, 2012). These approaches make predictions about how people generate of quantified descriptions (Briggs & Harner, 2019; Chen, van Deemter, & Lin, 2019). For instance, Briggs and Harner (2019) modeled the human data from Barr et al. (2013) through a method they called *perceptual cost pruning*, which removes particular symbolic information to mimic the perceptual costs required to encode particular pieces of information. The human data were best simulated by picking the most precise descriptions remaining after taking such costs into account. This procedure provides a computational implementation of the least effort hypothesis. The results of our study suggest that a better way of implementing the least effort hypothesis in computational systems is to instead design them to incrementally perceive and construct scene representations, thereby yielding the most human-like descriptions.

Acknowledgments

We would also like to thank Kevin Zish, Kalyan Gupta, and the Knexus Research Corporation for their assistance in supporting these studies. Additionally, we would like to thank the reviewers for helpful feedback in improving the presentation of this work. This work was supported by a NRC Research Associateship award to HH, a NRL Karles Fellowship awarded to GB, 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

Barr, D., van Deemter, K., & Fernández, R. (2013). Generation of quantified referring expressions: evidence from experimental data. In *Proceedings of the 14th european workshop on natural language generation* (pp. 157–161).

Briggs, G., & Harner, H. (2019). Generating Quantified Referring Expressions with Perceptual Cost Pruning. In *Proceedings of the 12th International Conference on Natural Language Generation* (pp. 11–18). Tokyo, Japan.

Briggs, G., Wasylyshyn, C., & Bello, P. F. (2019). Elicitation of Quantified Description Under Time Constraints. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society* (pp. 1436–1442). Montreal, Canada.

Burgess, A., & Barlow, H. (1983). The precision of numerosity discrimination in arrays of random dots. *Vision Research*, *23*(8), 811–820.

Cantrell, L., Kuwabara, M., & Smith, L. B. (2015). Set size and culture influence children's attention to number. *Journal of Experimental Child Psychology*, *131*, 19–37.

Cantrell, L., & Smith, L. B. (2013). Set size, individuation, and attention to shape. *Cognition*, *126*(2), 258–267.

Chen, G., van Deemter, K., & Lin, C. (2019). Generating quantified descriptions of abstract visual scenes. In *Proceedings of the 12th international conference on natural language generation* (pp. 529–539). Tokyo, Japan.

Ciccione, L., & Dehaene, S. (2020). Grouping mechanisms in numerosity perception. Retrieved from `https://doi.org/10.31234/osf.io/p6ryv`

Cousins, J., & Ginsburg, N. (1983). Subjective correlation and the regular-random numerosity illusion. *The Journal of general psychology*, *108*(1), 3–10.

Cummins, C. (2015). *Constraints on Numerical Expressions* (Vol. 5). Oxford University Press.

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition*, *44*(1-2), 1–42.

Fernberger, S. W. (1921). A preliminary study of the range of visual apprehension. *The American Journal of Psychology*, *32*(1), 121–133.

Frith, C. D., & Frit, U. (1972). The solitaire illusion: An illusion of numerosity. *Perception & Psychophysics*, *11*(6), 409–410.

Ginsburg, N. (1976). Effect of item arrangement on perceived numerosity: Randomness vs regularity. *Perceptual and motor skills*, *43*(2), 663–668.

Ginsburg, N. (1980). The regular-random numerosity illusion: Rectangular patterns. *The Journal of General Psychology*, *103*(2), 211–216.

Grice, H. P. (1975). Logic and conversation. *1975*, 41–58.

Hesse, C., & Benz, A. (2018). Giving the wrong impression: Strategic use of comparatively modified numerals in a question answering system. In *Proceedings of The Conference on Natural Language Processing (KONVENS)* (pp. 148–157). Vienna, Austria.

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

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

Krahmer, E., & Van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, *38*, 173–218.

Mandler, G., & Shebo, B. J. (1982). Subitizing: An analysis of its component processes. *Journal of Experimental Psychology: General*, *111*, 1–22.

Oberly, H. S. (1924). The range for visual attention, cognition and apprehension. *The American Journal of Psychology*, *35*(3), 332–352.

Okanda, M., Asada, K., Moriguchi, Y., & Itakura, S. (2015). Understanding violations of gricean maxims in preschoolers and adults. *Frontiers in psychology*, *6*, 901.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*, 411–419.

Poom, L., Lindskog, M., Winman, A., & Van den Berg, R. (2019). Grouping effects in numerosity perception under prolonged viewing conditions. *PloS one*, *14*(2).

Starkey, G. S., & McCandliss, B. D. (2014). The emergence of "groupitizing" in children's numerical cognition. *Journal of Experimental Child Psychology*, *126*, 120–137.

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

Van Deemter, K. (2016). *Computational models of referring: a study in cognitive science*. MIT Press.

Van Oeffelen, M. P., & Vos, P. G. (1982). Configurational effects on the enumeration of dots: Counting by groups. *Memory & Cognition*, *10*(4), 396–404.

Vos, P. G., Van Oeffelen, M. P., Tibosch, H. J., & Allik, J. (1988). Interactions between area and numerosity. *Psychological Research*, *50*(3), 148–154.