Be Precise or Fuzzy: Learning the Meaning of Cardinals and Quantifiers from Vision

Sandro Pezzelle[1], **Marco Marelli**[2] **and Raffaella Bernardi**[3]

[1]CIMeC - Center for Mind/Brain Sciences, University of Trento
[2]Department of Psychology, University of Ghent
[3]CIMeC/DISI, University of Trento
{sandro.pezzelle}@unitn.it
{marco.marelli}@ugent.be
{raffaella.bernardi}@unitn.it

Abstract

People can refer to quantities in a visual scene by using either exact cardinals (e.g. *one*, *two*, *three*) or natural language quantifiers (e.g. *few*, *most*, *all*). In humans, these two processes underlie fairly different cognitive and neural mechanisms. Inspired by this evidence, the present study proposes two models for learning the objective meaning of cardinals and quantifiers from visual scenes containing multiple objects. We show that a model capitalizing on a 'fuzzy' measure of similarity is effective for learning quantifiers, whereas the learning of exact cardinals is better accomplished when information about number is provided.

1 Introduction

In everyday life, people can refer to quantities by using either cardinals (e.g. *one*, *two*, *three*) or natural language quantifiers (e.g. *few*, *most*, *all*). Although they share a number of syntactic, semantic and pragmatic properties (Hurewitz et al., 2006), and they are both learned in a fairly stable order of acquisition across languages (Wynn, 1992; Katsos et al., 2016), these quantity expressions underlie fairly different cognitive and neural mechanisms. First, they are handled differently by the language acquisition system, with children recognizing their disparate characteristics since early development, even before becoming 'full-counters' (Hurewitz et al., 2006; Sarnecka and Gelman, 2004; Barner et al., 2009). Second, while the neural processing of cardinals relies on the brain region devoted to the representation of quantities, quantifiers rather elicit regions for general semantic processing (Wei et al., 2014). Intuitively, cardinals and quantifiers refer to quantities in a different way, with the former representing a mapping between a word and the exact cardinality of a set, the latter expressing a 'fuzzy' numerical concept denoting set relations

Figure 1: How many are *dogs*? Three/Most.

or proportions of sets (Barner et al., 2009). As a consequence, speakers can reliably answer questions involving quantifiers even in contexts that preclude counting (Pietroski et al., 2009), as well as children lacking exact cardinality concepts can understand and appropriately use quantifiers in grounded contexts (Halberda et al., 2008; Barner et al., 2009). That is, knowledge about (large) precise numbers is neither necessary nor sufficient for learning the meaning of quantifiers.

Inspired by this evidence, the present study proposes two computational models for learning the meaning of cardinals and quantifiers from visual scenes. Our hypothesis is that learning cardinals requires taking into account the number of instances of the target object in the scene (e.g. number of *dogs* in Figure 1). Learning quantifiers, instead, would be better accomplished by a model capitalizing on a measure evaluating the 'fuzzy' amount of target objects in the scene (e.g. proportion of 'dogness' in Figure 1). In particular, we focus on those cases where both quantification strategies might be used, namely scenes containing target (*dogs*) and distractor objects (*cats*). Our approach is thus different from salient objects detection, where the distinction targets/distractors is missing (Borji et al., 2015; Zhang et al., 2015; Zhang et al., 2016). With respect to cardinals, our approach is similar to (Seguí et al., 2015), who propose a model for counting people in natural

scenes, and to more recent work aimed at counting either everyday objects in natural images (Chattopadhyay et al., 2016) or geometrical objects with attributes in synthetic scenes (Johnson et al., 2016). With respect to quantifiers, our approach is similar to (Sorodoc et al., 2016), who use quantifiers *no*, *some*, and *all* to quantify over sets of colored dots. Differently from ours, however, all these works tackle the issue as either a classification problem or a Visual Question Answering task, with less focus on learning the meaning representation of each cardinal/quantifier. To our knowledge, this is the first attempt to jointly investigate both mechanisms and to obtain the meaning representaton of each cardinal/quantifier as resulting from a language-to-vision mapping.

Based on their geometric intepretation, we propose to use **cosine** and **dot product** similarity between the target object and the scene as our measures for quantifiers and cardinals, respectively. The former, ranging from -1 to 1, evaluates the similarity between two vectors with respect to their orientation and irrespectively of their magnitudes. That is, the more two vectors are overall similar, the closer they are. Ideally, cosine similarity between an image depicting a *dog* and a scene containing either 3 or 10 *dogs* without distractors (hence, 'all') should be equal to 1. Therefore, it would indicate that the proportion of 'dogness' in the scene is highest. Dot product, on the other hand, is defined as the product of the cosine between two vectors and their Euclidean magnitudes. By taking into account the magnitudes, this measure ideally encodes information regarding the number of times a target object is repeated in the scene. In the above-mentioned example, indeed, dot product would be 3 and 10, respectively. In this simplified setting, thus, it would be equal to the number of *dogs*.

Furthermore, we propose that the 'objective' meaning of each cardinal/quantifier can be learned by means of a cross-modal mapping (see Figure 4) between the linguistic representation of the target object and its quantity (either exact or fuzzy) in a visual scene. To test our hypotheses, we carry out a proof-of-concept on the synthetic datasets we describe in Section 2. First, we explore our visual data by means of the two proposed similarity measures (§ 3.1). Second, we learn the meaning representations of cardinals and quantifiers and evaluate them in the task of retrieving unseen combinations of targets/distractors (§ 3.2). As hypothesized, the two quantification mechanisms turn out to be better accounted for by models capitalizing on the expected similarity measures.

2 Data

In order to test our hypothesis, we need a dataset of visual scenes which crucially include multiple objects. Moreover, some objects in the scene should be repeated, so that we might say, for instance, that out of 5 objects 'three'/'most' are *dogs*. Although a large number of image datasets are currently available (see Lin et al. (2014) among many others), no one fully satisfies these requirements. Typically, images depict one salient object and even when multiple salient objects are present, only a handful of cases contain both targets and distractors (Zhang et al., 2015; Zhang et al., 2016). To bypass these issues, in the present work we experiment with synthetic visual scenes (hence, scenarios) that are made up by at most 9 images each representing one object. The choice of using a 'patchwork' of object-depicting images is motivated by the need of representing a reasonably large variability (e.g. 'few' refer to scenes containing 2 target objects out of 7 as well as 1/5, 4/9, etc.). This way, we avoid matching a quantifier always with the same number of target objects (except *no*, that is always represented by 0 targets), and allow cardinals to be represented by scenes with different numbers of distractors. At the same time, we get rid of any issues related to object localization.

We experiment with quantifiers (hence, Qs) *no*, *few*, *most*, and *all*, which we defined *a priori* by ratios 0%, 1-49%, 51-99% and 100%, respectively. Consistently with our goals, this arguably simplified setting does neither take into account pragmatic uses of Qs (i.e. we treat them as lying on an ordered scale) nor reflect possible overlappings. For these reasons, we avoid using quantifiers as *some* whose meaning overlaps with the meaning of many others. As far as cardinals (hence, Cs) are concerned, we experiment with scenarios in which the cardinality of the targets ranges from 1 to 4. Cs up to 4 are acquired by children incrementally at subsequent stages of their development, with higher numbers being learned upon this knowledge with the ability of counting (Barner et al., 2009). Also, Cs ranging from 1 to 3-4 are widely known to exhibit some peculiar properties

Train-q				Train-c			
no	few	most	all	one	two	three	four
0/1	1/6	2/3	1/1	1/1	2/2	3/3	4/4
0/2	2/5	3/4	2/2	1/3	2/3	3/4	4/5
0/3	2/7	3/5	3/3	1/4	2/5	3/5	4/6
0/4	3/8	4/5	4/4	1/6	2/7	3/8	4/7
Test-q				**Test-c**			
no	few	most	all	one	two	three	four
0/5	1/7	4/6	5/5	1/2	2/4	3/7	4/8
0/8	4/9	6/8	9/9	1/7	2/9	3/9	4/9

Table 1: Combinations in Train and Test.

(i.e. their exact number can be immediately and effortlessly grasped) due to which they are usually referred to as 'subitizing' range (Piazza et al., 2011; Railo et al., 2016).

2.1 Building the scenarios

We use images from ImageNet (Deng et al., 2009). Starting from the full list of 203 concepts and corresponding images extracted by Cassani (2014), we discarded those concepts whose corresponding word had low/null frequency in the large corpus used in (Baroni et al., 2014). To get rid of issues related to concept identification, we used a single representation for each of the 188 selected concepts. Technically, we computed a centroid vector by averaging the 4096-dimension visual features of the corresponding images, which were extracted from the *fc7* of a CNN (Simonyan and Zisserman, 2014). We used the VGG-19 model pretrained on the ImageNet ILSVRC data (Russakovsky et al., 2015) implemented in the MatConvNet toolbox (Vedaldi and Lenc, 2015). Centroid vectors were reduced to 100-d via PCA and further normalized to length 1 before being used to build the scenarios. When building the scenarios, we put the constraint that distractors have to be different from each other. Moreover, only distractors whose visual cosine similarity with respect to the target is lower than the average are selected. For each scenario, target and distractor vectors are summed together. As a result, each scenario is represented by a 100-d vector.

We also experimented with scenarios where vectors are concatenated to obtain a 900-d vector (empty 'cells' are filled with 0s vectors) and further reduced to 100-d via PCA. Since the pattern of results in the only-vision evaluation (see § 3.1) turned out to be similar to the results obtained in the 'summed' setting, due to space limitations we will only focus on the 'summed' setting.

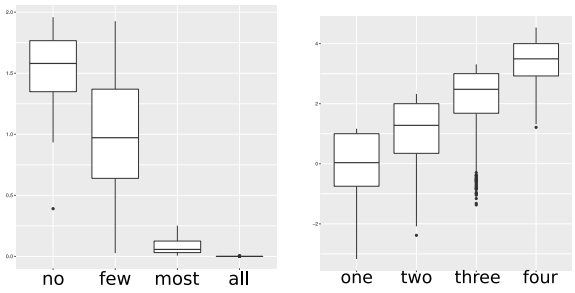

Figure 2: Left: quantifiers against cosine distance. Right: cardinals against dot product.

2.2 Datasets

We built one dataset for Cs and one for Qs, each containing 4512 scenarios.[1] We then split each of the two in one 3008-datapoint Training Dataset (**Train**) for training and validation and one 1504-datapoint Testing Dataset (**Test**) for testing. The two datasets were split according to their 'combinations', that is the mixture of targets and distractors in the scenario. As reported in Table 1, we kept 4 different combinations for each C/Q in Train and 2 in Test. Note that the numerator refers to the number of targets, the denominator to the total number of objects. The number of distractors is thus given by the difference between the two values. To illustrate, in Train-q 'few' is represented by scenarios 1/6, 2/5, 2/7, and 3/8, whereas in Test-q 'few' is represented by scenarios 1/7 and 4/9. The initial 4512 scenarios have been obtained by building a total of 24 different scenarios (6 combinations * 4 C/Q classes) for each of the 188 objects. A particular effort has been paid in making the datasets as balanced as possible. When designing the combinations for 'few' and 'most', for example, we controlled for the proportion of targets in the scene, in order to avoid making one of the two easier to learn. Also, combinations were thought to avoid biasing cardinals toward fixed proportions of targets/distractors.

3 Experiments

3.1 Only-vision evaluation

As a first step, we carry out a preliminary evaluation aimed at exploring our visual data. If our intuition about the information encoded by the two similarity measures is correct (see § 1), we

[1] A visual representation of our scenarios is provided in the rightmost side of Figure 4, while Figure 1 is only intended to provide a more intuitive overview of the task.

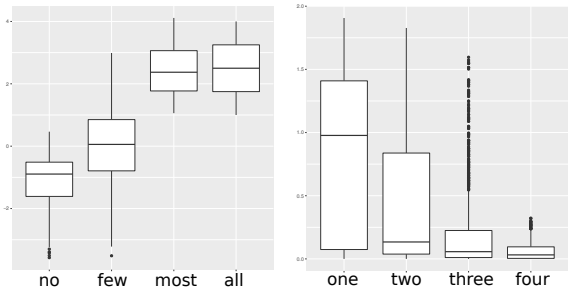

Figure 3: Left: quantifiers against dot product. Right: cardinals against cosine distance.

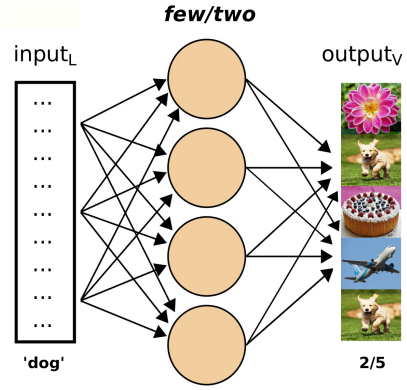

Figure 4: One learning event of our proposed cross-modal mapping. Cosine is used for quantifiers (*few*), dot product for cardinals (*two*).

should observe that cosine is more effective than dot product in distinguishing between different Qs, while the latter should be better than cosine for Cs. Moreover, Qs/Cs should lie on an ordered scale. To test our hypothesis, we compute cosine distances (i.e. 1−cosine, to avoid negative values) and dot product similarity for each target-scenario pair in both Train and Test (e.g. *dog* vs 2/5 *dogs*). Figure 2 reports the distribution of Qs with respect to cosine (left) and Cs with respect to dot product (right) in Train. As can be seen from the boxplots, both Qs and Cs are ordered on a scale. In particular, cosine distance is highest in *no* scenarios (where the target is not present), lowest in *all* scenarios. For Cs, dot product is highest in *four* scenarios, lowest in *one* scenarios.

Our intuition is further confirmed by the results of a radial-kernel SVM classifier fed with either cosine or dot product similarities as predictors.[2] Qs are better predicted by cosine than dot product (78.6% vs 63.8%), whereas dot product is a better predictor of Cs than cosine (68.7% vs 44.7%). As shown in Figure 3, the ordered scale is indeed represented to a much lesser extent when Qs are plotted against dot product (left) and Cs against cosine (right). A similar pattern of SVM results and similar plots emerged when experimenting with Test.

3.2 Cross-modal mapping

Our core proposal is that the meaning of each C/Q can be learned by means of a cross-modal mapping between the linguistic representation of the target object (e.g. *dog*, *mug*, etc.) and a number of scenarios representing the target object in a given C/Q setting (e.g. 'two'/'few' *dogs*). In our approach, each word (e.g. *dog*) is represented by a 400-d embedding built with the `CBOW` architecture of `word2vec` (Mikolov et al., 2013) and the best-predictive parameters of Baroni et al. (2014) on a 2.8B tokens corpus. The original 400-d vectors are further reduced to 100-d via PCA before being fed into the model.

Figure 4 reports a single learning event of our proposed model. Each C/Q (e.g. *two*, *few*) is learned as a separate function that maps each of the 188 words representing our selected concepts to its corresponding 4 scenarios in Train (see § 2.2). To illustrate, the meaning of *few* is learned by mapping each word into the 4 visual scenes where the amount of 'targetness' is less than 50% (see § 2), whereas *two* is learned by mapping each word to the scenarios where the number of targets is 2, and so on. This mapping, we conjecture, would mimic the multimodal mechanism by which children acquire the meaning of both Cs and Qs (see Halberda et al. (2008)). Once learned, the function representing each C/Q can be evaluated against scenarios containing an unseen mixture of (known) target objects and distractors. If it has encoded the correct meaning of the quantified expression, the function will retrieve the unseen scenarios containing the correct quantity (either exact or fuzzy) of target objects.

We experiment with three different models: linear (**lin**), cosine neural network (**nn-cos**), dot-product neural network (**nn-dot**). The first model is a simple linear mapping. The second is a single-layer neural network (activation function ReLU) that maximizes the cosine similarity between input (linguistic) and output vector (visual). The third is a similar neural network that approximates to 1 the

[2] We experimented with linear, polynomial, and radial kernels. We only report results obtained with default radial kernel, that turned out to be the overall best model.

	lin		nn-cos		nn-dot	
	mAP	*P2*	*mAP*	*P2*	*mAP*	*P2*
no	0.78	0.65	**0.87**	0.77	0.54	0.37
few	0.59	0.39	**0.68**	0.51	0.59	0.43
most	0.61	0.36	0.60	0.29	**0.62**	0.45
all	0.75	0.66	**1**	1	0.33	0.12
one	0.44	0.30	0.38	0.21	**0.61**	0.45
two	0.35	0.15	0.38	0.21	**0.57**	0.43
three	0.38	0.16	0.36	0.13	**0.56**	0.40
four	0.65	0.47	0.75	0.60	**0.76**	0.61

Table 2: R-target. *mAP* and *P2* for each model.

	no	few	most	all
no	**288**	88	0	0
few	141	**191**	38	6
most	0	0	111	265
all	0	0	0	**376**
	one	two	three	four
one	**168**	113	54	41
two	64	**136**	124	52
three	23	80	130	145
four	10	24	72	**272**

Table 3: Top: Q nn-cos, number of cases retrieved in top-2 positions. Bottom: same for C nn-dot.

dot product between input and output. We evaluate the mapping functions by means of a retrieval task aimed at picking up the correct scenarios from Test among the set of 8 scenarios built upon the same target object. Recall that in Test there are 2 combinations * 4 C/Q classes for each concept.

Results As reported in Table 2, nn-cos is overall the best model for Qs, whereas nn-dot is the best model for Cs. In particular, mean average precision (*mAP*) is higher in nn-cos for 3 out of 4 Qs, with only *most* reaching slightly better mAP in Q nn-dot due to the high number of cases confounded with *all* by the Q nn-cos model (see Table 3). Conversely, both mAP and precision at top-2 positions (*P2*) for Cs are always higher in nn-dot compared to the other models. From a qualitative analysis of the results, it emerges that both the best-predictive models make 'plausible' errors, i.e. they confound Cs/Qs that are close to each other in the ordered scale. Table 3 reports the confusion matrices for the best performing models. Besides retrieving more cases of *all* instead of (correct) *most*, the Q nn-cos model often confounds *few* with *no*. Similarly, the C nn-dot model often confounds *three* with *four*, *one* with *two*, *two* with *three*, and so on. Overall, both models pick up very few or no responses that are on the opposite end of the 'scale', thus suggesting that the meaning representation they learn encodes, to a certain extent, information about the ordered position of the quantified expressions.

4 Discussion

We propose that the meaning of Cs and Qs can be learned by means of a language-to-vision mapping, and we show that two models capitalizing on dot product and cosine better account for Cs and Qs, respectively. In future research, we plan to further investigate this issue by using real-scene images to avoid constraining the visual data. Moreover, we plan to experiment with a broader set of quantifiers (e.g. *some*, *almost all*, etc.) and higher cardinals. The latter investigation, in particular, would allow us to verify whether our approach is suitable for the (potentially infinite) set of 'cardinal functions' beyond the subitizing range. If so, we might observe that the models keep making cognitively plausible errors, picking items that are close to the target one in the ordered scale. This evidence, we believe, would further motivate our 'one quantifed expression, one function' approach, which is partially inspired by the evidence that, in human brain, so-called number neurons are tuned to preferred numbers (Nieder, 2016). Simplifying somewhat, each number would activate specific neurons. Finally, we believe that taking into account speakers' uses of Cs and Qs would constitute the natural next step toward a complete modelling of the meaning of quantified expressions.

Acknowledgments

We are very grateful to Germán Kruszewski for the invaluable contribution in developing and discussing the intuitions behind this work. We are also grateful to Marco Baroni, Aurélie Herbelot, Gemma Boleda and Ravi Shekhar for their advice and feedback. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPUs used in our research, and the iV&L Net (ICT COST Action IC1307) for funding the second author's research visit aimed at working on this project.

References

David Barner, Amanda Libenson, Pierina Cheung, and Mayu Takasaki. 2009. Cross-linguistic relations between quantifiers and numerals in language acquisition: Evidence from Japanese. *Journal of experimental child psychology*, 103(4):421–440.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)*, pages 238–247.

Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. 2015. Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, 24(12):5706–5722.

Giovanni Cassani. 2014. Distributional semantics for child directed speech: A multimodal approach. Master's thesis, University of Trento.

Prithvijit Chattopadhyay, Ramakrishna Vedantam, Ramprasaath RS, Dhruv Batra, and Devi Parikh. 2016. Counting everyday objects in everyday scenes. *arXiv preprint arXiv:1604.03505*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE.

Justin Halberda, Len Taing, and Jeffrey Lidz. 2008. The development of 'most' comprehension and its potential dependence on counting ability in preschoolers. *Language Learning and Development*, 4(2):99–121.

Felicia Hurewitz, Anna Papafragou, Lila Gleitman, and Rochel Gelman. 2006. Asymmetries in the acquisition of numbers and quantifiers. *Language learning and development*, 2(2):77–96.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2016. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *arXiv preprint arXiv:1612.06890*.

Napoleon Katsos, Chris Cummins, Maria-José Ezeizabarrena, Anna Gavarró, Jelena Kuvač Kraljević, Gordana Hrzica, Kleanthes K Grohmann, Athina Skordi, Kristine Jensen de López, Lone Sundahl, et al. 2016. Cross-linguistic patterns in the acquisition of quantifiers. *Proceedings of the National Academy of Sciences*, 113(33):9244–9249.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014*, pages 740–755. Springer.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Andreas Nieder. 2016. The neuronal code for number. *Nature Reviews Neuroscience*.

Manuela Piazza, Antonia Fumarola, Alessandro Chinello, and David Melcher. 2011. Subitizing reflects visuo-spatial object individuation capacity. *Cognition*, 121(1):147–153.

Paul Pietroski, Jeffrey Lidz, Tim Hunter, and Justin Halberda. 2009. The meaning of 'most': Semantics, numerosity and psychology. *Mind & Language*, 24(5):554–585.

Henry Railo, Veli-Matti Karhu, Jeremy Mast, Henri Pesonen, and Mika Koivisto. 2016. Rapid and accurate processing of multiple objects in briefly presented scenes. *Journal of vision*, 16(3):8–8.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

Barbara W Sarnecka and Susan A Gelman. 2004. Six does not just mean a lot: Preschoolers see number words as specific. *Cognition*, 92(3):329–352.

Santi Seguí, Oriol Pujol, and Jordi Vitria. 2015. Learning to count with deep object features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 90–96.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Ionut Sorodoc, Angeliki Lazaridou, Gemma Boleda, Aurélie Herbelot, Sandro Pezzelle, and Raffaella Bernardi. 2016. 'Look, some green circles!': Learning to quantify from images. In *Proceedings of the 5th Workshop on Vision and Language at ACL*.

Andrea Vedaldi and Karel Lenc. 2015. *MatConvNet – Convolutional Neural Networks for MATLAB*. Proceeding of the ACM Int. Conf. on Multimedia.

Wei Wei, Chuansheng Chen, Tao Yang, Han Zhang, and Xinlin Zhou. 2014. Dissociated neural correlates of quantity processing of quantifiers, numbers, and numerosities. *Human brain mapping*, 35(2):444–454.

Karen Wynn. 1992. Children's acquisition of the number words and the counting system. *Cognitive psychology*, 24(2):220–251.

Jianming Zhang, Shugao Ma, Mehrnoosh Sameki, Stan Sclaroff, Margrit Betke, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech. 2015. Salient object subitizing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4045–4054.

Jianming Zhang, Shuga Ma, Mehrnoosh Sameki, Stan Sclaroff, Margrit Betke, Zhe Lin, Xiaohui Shen, Brian Price, and Radomír Měch. 2016. Salient object subitizing. *arXiv preprint arXiv:1607.07525*.