Probing the Mental Representation of Quantifiers

Sandro Pezzelle[a,*], Raffaella Bernardi[a,b], Manuela Piazza[a]

[a] *CIMeC - Center for Mind/Brain Sciences, University of Trento, Corso Bettini, 31, Rovereto, Italy*
[b] *DISI, University of Trento, Via Sommarive, 9, Trento, Italy*

Abstract

In this study, we investigate the mental representation of non-numerical quantifiers ("some", "many", "all", etc.) by comparing their use in *abstract* and in *grounded* perceptual contexts. Using an approach similar to that used in the number domain, we test whether (and to what extent) such representation is constrained by the way we perceive the world through our senses. In two experiments, subjects either judged the similarity of quantifier pairs (presented as written words) or chose among a predetermined list of quantifiers the one that best described a visual image depicting a variable number of target and non-target items. The results were rather consistent across experiments, and indicated that quantifiers are mentally organized on an ordered but non-linear compressed scale where the quantifiers that imply small quantities appear more precisely differentiated across each other compared to those implying large quantities. This fits nicely with the idea that we construct our representations of such symbols mainly by mapping them to the representations of quantities that we derive from perception.

Keywords: quantifiers, proportions, quantity estimation, subitizing, grounding, scale

1. Introduction

One of the common goals of linguists and cognitive scientists is to uncover and formally characterize how linguistic symbols are mentally represented. Here we attack the issue by focusing on a specific class of words, that of quantifiers (words like "some", "many", "few", "a lot", "all", "none").

[*] Corresponding author
Email address: sandro.pezzelle@unitn.it (Sandro Pezzelle)

Quantifiers have long been considered as a particularly intriguing class of words especially by linguists, since they display several peculiar properties. First, from a formal semantic perspective they are conceived as non-referential (Montague, 1973; Barwise & Cooper, 1981; Westerståhl, 1985; Van Benthem et al., 1986; Keenan & Stavi, 1986; Szabolcsi, 2010): Differently from many other words, quantifiers do not denote objects, but instead relations between sets of objects. Second, quantifiers are widely affected by the linguistic context of use. This particularly holds for some quantifiers, like "few" and "many", which have therefore been proposed to be non-extensional (Keenan & Stavi, 1986; Westerståhl, 1985): The two sentences "Many doctors attended the meeting this year" and "Many lawyers attended the meeting this year" (even assuming that the doctors and lawyers attending the respective meetings are equal in number) might have different truth values depending on the number of doctors and lawyers who used to attend the meeting. Third, from a pragmatic perspective it has been shown how the different degree of information or logical strength of the quantifiers (that "some" is less informative than "all") affects the implicit information that people *infer* from an utterance (Horn, 1984). For example, listening to the sentence "Some students were satisfied with the marks" a hearer would infer that "Not all the students were satisifed". Fourth, quantifiers cannot be simplistically considered as words that stand for amounts, numbers, proportions (Moxey & Sanford, 1993, 2000; Paterson et al., 2009; Nouwen, 2010). Even when expressing approximately the same quantity (e.g. "few" and "a few"), quantifiers differ from each other with respect to the perspective they give to this quantity, by bringing the hearer to focus on either the target set ("a few") or the non-target set ("few"). For instance, "few of these cars break down" is likely to bring the hearer's attention to the vast majority of cars that do not break down. "A few of these cars break down", instead, is more likely to bring the attention to the cars that do break. This difference in the focus influences the hearer's behavior in a positive/negative way (Moxey & Sanford, 2000; Paterson et al., 2009). Consequently, quantifiers have been described in terms of probability distributions over scales (Moxey & Sanford, 1993; Yildirim et al., 2013; Schöller & Franke, 2017). Finally, the variability of quantifiers across conditions, together with their rather elusive status with respect to the traditional linguistic classifications, have brought some researchers to take the extreme stance that devising a general semantics for these expressions might not even be possible (Nouwen, 2010).

Although a long tradition of studies convincingly proved that numerical information, such as the mechanisms of quantity estimation and comparison, is fundamental in the comprehension of quantifiers (Heim et al., 2012; Shikhare et al., 2015; Deschamps et al., 2015),[1] cognitive science has not been successful at characterizing how humans mentally represent quantifiers. Historically, even if there has been a shared intuitive assumption that quantifiers might be internally represented on an ordered scale (which some conceived as governed by absolute quantities, e.g. Newstead et al. (1987), and other by proportions, e.g. Graves & Hodge (1943); Hammerton (1976)), there has been little attempt at formally trying to capture the features of such scale in a quantitative manner. One approach has been to investigate the conditions of the external world that trigger the use of the different quantifiers: Subjects, presented with sets of a various number of target and non-target (visual) items, are asked either to pick, among a predetermined list, the quantifier that best fits the scene or to rate the appropriateness of a list of scene-quantifier associations. Studies of this sort are only very few, and they are hard to compare as they each investigate different sets of quantifiers, as well as slightly different aspects of the stimuli (some analyze the effect of the number of targets, e.g. Newstead & Coventry (2000), some the number of both targets and non-targets, e.g. Coventry et al. (2005, 2010), some the proportion of targets in the scene, e.g. Oaksford et al. (2002), often taking into account perceptual factors like the size of the items, their spatial arrangements or their category, e.g. Newstead & Coventry (2000); Coventry et al. (2010)), though without investigating the potential interactions across all the possible variables. Moreover, the experimental design of all these studies lacks cases where the various effects can be disentangled, for example visual scenes with a small number of targets corresponding to a high proportion (e.g., 3 targets out of 4 total objects). Although with some inconsistencies, the results of these studies overall suggest that quantifiers are evaluated by taking into account the number of both targets and non-targets such that, given a fixed number of non-targets, scenarios with increasing targets are associated with

[1] This work typically employs a verification task: Given a scene depicting a variable proportion of target and non-target dots and a sentence embedding a quantified expression, participants are asked to quickly verify the semantic truth value of the sentence. What these studies showed is that errors and reaction times are typically affected by perceptual difficulty in observance to Weber's law.

quantifiers implying “larger” quantities. A notable exception is that, when the targets are very few, the number of non-targets seems not to play a role (Coventry et al., 2005). This indirectly suggests that quantifiers might be represented on an internal scale based on proportions which behaves somewhat differently for small sets. What these studies lack, however, is a quantitative characterization of the laws subtending the relation between quantifiers and perceptual stimulation and thus a thorough description of the internal scale.

Another complementary approach that psychologists have used to infer the structure of mental representations is that of directly asking subjects to compare words pairwise and to rate, on a given scale, their semantic similarity in a purely linguistic context (with no direct relation to concrete objects/sets). This way, the potential confounds due to the constraints imposed by perception are eliminated. In this approach, the analysis of the global pattern of rated distances across words can then be used to reconstruct the internal geometry of the representational space of those words (using Multi-Dimensional Scaling, e.g. Arnold (1971); Steyvers et al. (2004)). To our knowledge, this approach has been applied to the domain of quantifiers only by Holyoak & Glass (1978), who experimented with a set of five items. Studies of this sort would be crucial for complementing the studies that explore quantifiers in grounded conditions. In particular, the comparison across the grounded and abstract use of quantifiers is useful to approach the question of to what extent the mental representations of quantifiers (and, more generally, of symbols) are, or are not, constrained by the way we perceptually elaborate the objects or objects features to which the symbols are typically used to refer to.

While the abstract view of semantics predicts that symbols are mainly organized according to purely linguistic variables (frequency of use, frequency of association in the lexicon, antinomy, etc.), the grounded cognition view predicts that symbols are mentally represented in a way that at least partially reflects (or is isomorphic to) the way we perceive the world through our senses. This should be reflected both in how subjects use quantifiers to describe perceptual scenes, and in purely abstract contexts when they evaluate quantifiers among each other. This approach has been taken for example in the number domain, where several pieces of data indicate that the internal representation of number symbols (words or Arabic digits, denoting cardinals) appears as governed by the same representational constraints that govern the perception of numerosities in concrete sets, namely on an internal scale which appears overall logarithmically compressed (see Piazza & Eger (2016), for a recent review). This is the case both when number symbols are compared among each other and when they are used to describe perceptual scenes (e.g. Izard & Dehaene (2008)). The aim of the current paper is to export this approach to study the mental space of quantifiers, its main dimensions, and its internal geometry, and to contrast the predictions from the abstract cognition and the grounded cognition comparing grounded-perceptual and abstract tasks: Using a common list of quantifiers and two large groups of subjects, one experiment investigates quantifiers in grounded conditions, asking subjects to describe visual scenes choosing the most appropriate quantifier (Experiment 1), and the other investigates quantifiers in a purely linguistic context, asking subjects to rate the similarity among quantifier word pairs (Experiment 2).

2. Methods

Two experiments were administered to native-Italian participants and employed the same set of 9 Italian quantifiers. The quantifiers used were *nessuno* ("none"), *quasi nessuno* ("almost none"), *la minor parte* ("the smaller part"), *pochi* ("few"), *alcuni* ("some"), *molti* ("many"), *la maggior parte* ("most"), *quasi tutti* ("almost all"), *tutti* ("all"). For sake of clarity, English translations will be used from now on throughout the paper. The selection of the quantifiers was aimed at experimenting with a fairly comprehensive set, including logical-Aristotelian ("none", "some", "all"), proportional ("the smaller part", "most"), and a range of other common quantifiers ("few", "many", "almost none", "almost all"). Moreover, an equal number of low-magnitude ("none", "almost none", "few", "the smaller part") and high-magnitude quantifiers ("many", "most", "almost all", "all") was ensured. Note that we did not consider "some" as belonging *a priori* to one or the other group.

2.1. Grounded task: Quantifiers used in perception

Thirty native-Italian participants (21 females, 9 males) with normal or corrected-to-normal vision carried out the task of evaluating 340 synthetic visual scenes containing

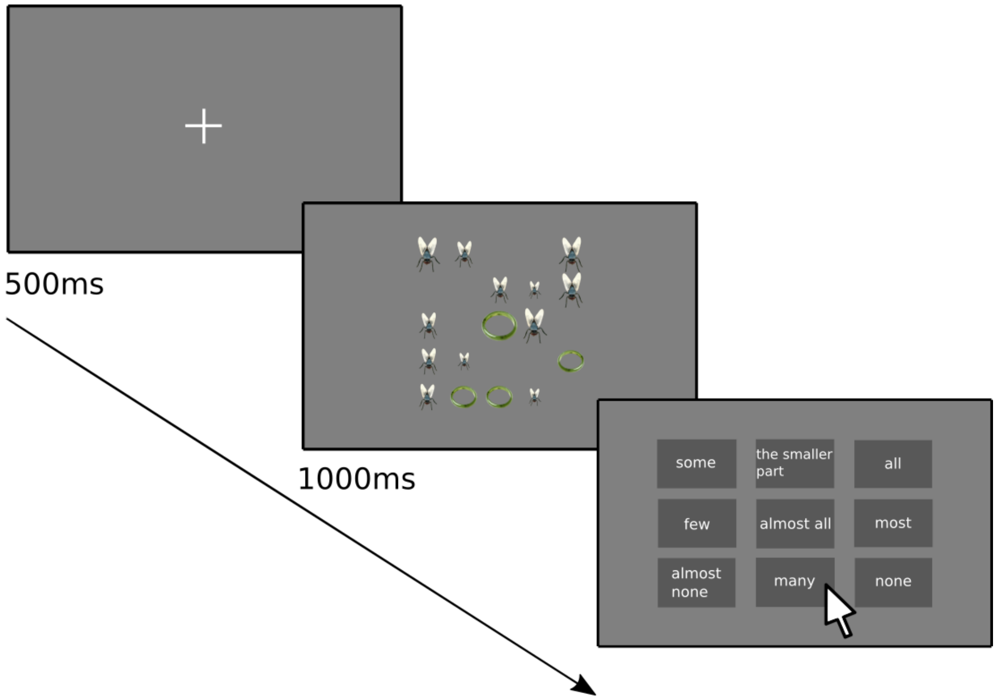

Figure 1: Schematic representation of the experiment. After a fixation cross of 500ms, a trial is presented for 1,000ms. Then the participant is asked to click on the quantifier that better describes the scene.

two categories of objects: Animals and artifacts. The total number of objects in the scene ranged from 3 to 20 (see section 2.1.1 for a detailed description of the visual stimuli), and the number of items in each of the two categories varied from 0 to 20. The experiment was implemented in Matlab using the Psychtoolbox-3 package. All participants performed the experiment in a quiet, dimly lit room at the CIMeC Psychophysic lab (Rovereto, Italy) using the same desktop computer, same monitor (size 23.6", resolution 1920x1080 pixels), and same mouse, and sitting at a distance of approximately 50cm from the screen. Eighteen participants requested and obtained university credits for their participation.

Before starting, two instruction pages describing the task were displayed. Participants were asked to be as accurate and fast as possible. The task consisted of attending the visual scene and to select the quantifier which better answered the question: "How many of the objects are animals?". Particular focus was put on the fact that the quantifier had to be chosen *always* with respect to the set of *animals* (target set). This choice was aimed at diminishing the chance of errors merely due to wrong associations between the question and the target set. By fixing the set of animals as the target set, in fact, participants should be more focused on the quantification task *per se*. Importantly, the 9 quantifiers were never presented in any kind of order during the instructions.

After reading the instructions and having clarified any possible doubt with the experimenter, a training session was provided to get familiar with the task. The training session comprised of 5 trials which were not included in the 340 test stimuli. The procedure was the same as the test session (see Figure 1 for a schematic representation of the experiment): A white fixation cross was presented for 500ms in the center of a grey background screen; afterwards, a visual scene was displayed for $1,000$ms followed by the 9 quantifiers presented in a 3*3-cell grid centered in the middle of the screen. The cells were well-spaced to prevent unwanted clicks, and highlighted by a darker shade of grey. Importantly, quantifiers were presented at each trial in a randomized position to avoid any familiarization effects. The task was to click on the chosen quantifier in the shortest possible time. After the response, a fixation cross appeared for 500ms followed by the next stimulus. After the first 5 training trials, a display was presented offering the possibility to train for extra 5 trials, different from the previous ones and also not included in the test set. Participants were asked to choose between training more or moving to the test session.

Before starting the test session, an instruction page was presented to specify that the experiment comprised of 10 blocks of 34 stimuli each. Subjects were reminded of the task. After left-clicking the mouse, participants started the first block of the experiment. At the end of each block, participants were allowed to take a self-paced pause. On each trial we recorded the chosen quantifier, its position on the grid, and the time taken to give the response. For each trial we also recorded a number of perceptual features describing the visual scene, such as the cardinality of animals and artifacts, their size (small, medium, large), and the ratio between animals and artifacts.

Responses by all participants were retained. 15 participants were in the age range 18-23, 11 in the range 24-29, 4 in the range 30-36. Seventeen requested and obtained university credits for their participation.

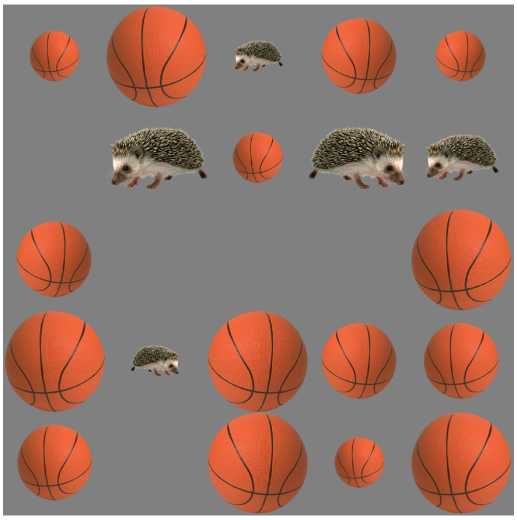

Figure 2: One visual scene used in the experiment, representing a targets:non-targets ratio of 1:3 (i.e. 25% of total items are targets).

2.1.1. Materials

The visual scenes used in the experiment consisted in multiple colored pictures of animals (hence, targets) and artifacts (hence, non-targets) displayed on the top of a grey background (see Figure 2). Scenes differed on the total number of items displayed, that could vary from 3 to 20. Across scenes, the number of targets and non-targets varied such that different targets:non-targets *ratios* were equally represented. Crucially, each ratio corresponded to a fixed proportion of targets with respect to the total number of objects (i.e., targets+non-targets) in the scene. For example, ratio 1:3 corresponded to 25% of targets (see Figure 2). We used 17 ratios, each presented 20 times during the experiment, out of which 8 were "positive" (targets $>$ 50%), 8 "negative" (targets $<$ 50%) and 1 "parity" (targets = 50%). Because each ratio could be generated by different combinations of cardinalities (e.g., ratio 1:4 could result from the combination of 1 target and 4 non-targets, as well as 2 targets and 8 non-targets, etc.), for each ratio we presented all possible combinations of cardinalities. For any possible combination, a fixed number of visual scenes was built.

Visual scenes were generated with an inhouse Matlab script using the following pipeline: Two pictures, one depicting a target (e.g. an instance of a hedgehog) and one depicting a non-target (e.g. an istance of a basketball) were randomly chosen from a sample of the database by Kiani et al. (2007) including 100 instances of targets and 145 instances of non-targets. The sample was previously obtained by manually selecting pictures depicting whole items (not just parts) and whose color, orientation, and shape were not deceptive (for example, we discarded pictures depicting butterfly-shaped pasta as their target/non-target categorization could have been problematic). The target and the non-target pictures were randomly inserted by the script onto a 5*5-cell virtual grid. In order to inject some variability, each picture was randomly assigned to one orientation on the vertical axis (right or left) and one size (large, medium, small size, corresponding to approximately 5.3°, 3.4°, and 2.3° of visual angle). None of the scenes contained objects that were all the same size. As for the orientation, its effect is less measurable since it depends on the visual properties of the object (see, e.g., the different effects on the hedgehog and the basketball in Figure 2). However, this is not an issue since we are not interested in formally investigating the role of object orientation in the task. In total, 340 visual scenes were included in the experiment, together with additional 10 trials for training.

2.2. Abstract task: Semantic similarity judgements

Thirty-three native-Italian participants (10 males, 22 females, 1 n.d.) completed this task. In an online survey powered by Google Forms, they were presented with pairs of quantifiers (e.g., "almost none" and "none"), and asked to rate their semantic similarity using a 7-point Likert-like scale, where 1 meant "highly dissimilar" and 7 "highly similar". Before starting the task, participants were presented with an instruction page where the terminology was briefly explained and the task exemplified. They were instructed that, in cases of difficulties in assessing the degree of semantic similarity between two quantifiers, they could adopt the strategy of mentally placing them into a default sentence (e.g., "*Few/Many* students have had high marks"), and judging the semantic similarity of the two resulting sentences. In order not to bias participants, only two trivial examples were provided in the instructions, namely "all-none"=1, and "some-some"=7. Moreover, given the constrained number of combinations, i.e. 9*9=81, no trial items were included. Each participant was asked to judge all 81 possible combinations in a randomized order of presentation. Each quantifier pair was rated twice by each participant, once in one order (i.e. “all-none”) and once in the opposite order (i.e. “none-all”). To avoid any priming or repetition bias, we ensured that the two versions of the same pair never occurred in a row. Even though no time limits were set, participants were asked to provide their judgements as accurately as possible in the shortest possible time.

One participant’s responses were discarded due to the repeated choice of the judgement 1 (i.e. “highly dissimilar”) in 55 out of 81 cases (68%). Responses by thirty-two participants (9 males, 22 females, 1 n.d.) were retained. 13 participants were in the age range 18-23, 14 in the range 24-29, 3 in the range 30-36, 2 in the range 37-42. Fifteen requested and obtained university credits for their participation.

3. Analysis and results

3.1. Grounded task: Quantifiers used in perception

All 30 participants successfully completed the experiment and provided each 340 responses. In total, $10,200$ datapoints were collected. To ensure the quality of the responses, we removed those datapoints for which the reaction times exceeded the average of 2.5 SD. We did not perform any other filtering of the data. In total, 257 responses were discarded, equal to 2.52% of total. All statistical analyses were performed in the R environment on the resulting sample. For each quantifier, in Table 1 we report the following descriptive statistics: (a) The total number of responses assigned, (b) the average proportion of targets out of total number of items, (c) the average number of targets, (d) the average number of non-targets, (e) the average total number of items. Note that quantifiers are sorted according to ascending (b), which also corresponds to ascending (c).

As can be seen in the table, “most” is the most used quantifier with $2,110$ responses. Low-magnitude quantifiers (“none”, “almost none”, “few”, “the smaller part”) are used $3,841$ times (38.6%), high-magnitude quantifiers (“all”, “almost all”, “many”, “most”) $4,706$ times (47.3%). As far as both the proportion and the cardinality of targets are concerned, the quantifiers turn out to be ordered on the following scale: “none”, “almost none”, “few”, “the smaller part”, “some”, “many”, “most”, “almost all”, “all”. By looking at the proportions defining each quantifier, an almost perfect mirroring can be observed between “none-all” ($\sim$ 0%-100%), “almost none-almost all” ($\sim$ 20%-80%), “the

quantifier	**(a) resp**	**(b) % targ**	**(c) n targ**	**(d) n non-targ**	**(e) n total**
none	604	0.01 (0.09)	0.13 (1.01)	11.35 (5.04)	11.48 (4.93)
almost none	861	0.19 (0.13)	1.69 (1.95)	7.81 (4.67)	9.45 (5.12)
few	1241	0.26 (0.13)	2.92 (1.58)	9.63 (4.96)	12.55 (5.40)
the smaller part	1135	0.32 (0.13)	3.79 (2.01)	8.99 (4.56)	12.78 (5.26)
some	1396	0.44 (0.13)	4.97 (2.30)	6.82 (3.66)	11.79 (4.79)
many	770	0.64 (0.14)	8.75 (3.76)	4.89 (2.66)	13.65 (4.53)
most	2110	0.69 (0.13)	8.82 (4.21)	3.90 (2.30)	12.72 (5.03)
almost all	1222	0.80 (0.12)	9.38 (5.08)	2.24 (2.00)	11.62 (5.68)
all	604	0.99 (0.09)	11.31 (5.04)	0.15 (1.13)	11.47 (4.99)

Table 1: Descriptive statistics. Columns are sorted with respect to ascending proportion of targets (b), which also corresponds to ascending cardinality of targets (c). Values in brackets refer to SD.

smaller part-most“ ($\sim$ 30%-70%). Such a pattern can be better observed in Figure 3, which shows the frequency distribution of responses across proportions of targets. As can be seen, the quantifiers involved in these pairs have similar “peaks” and distributions, though different frequencies.

In order to explore the role of cardinality of the target items in the scene, we separated the trials where the target items fell within the range of extremely well enumerable cardinalities (i.e. the so called “subitizing” range, corresponding to scenes containing up to 3 animals) from those containing more than 3 items. The distribution of responses can be observed in Figure 4, which reports quantifiers frequency for scenes within the subitizing range (leftmost panel) and exceeding the subitizing range (rightmost panel). It should be noted that while in the former the whole range of quantifiers is used (though “many” has an extremely low frequency), in the latter both “none” and “almost none” disappear, with an increasing use of quantifiers like “most” and “many”. It is worth mentioning that the choice of setting the subitizing threshold to 3 was aimed at making our results directly comparable to those reported by Coventry et al. (2005, 2010), who experimented with such setting.

To more formally investigate which factors contribute in determining quantifiers meaning in grounded contexts, we performed statistical analyses on the collected data. Because our variables of interest are naturally highly correlated (crucially, proportion of

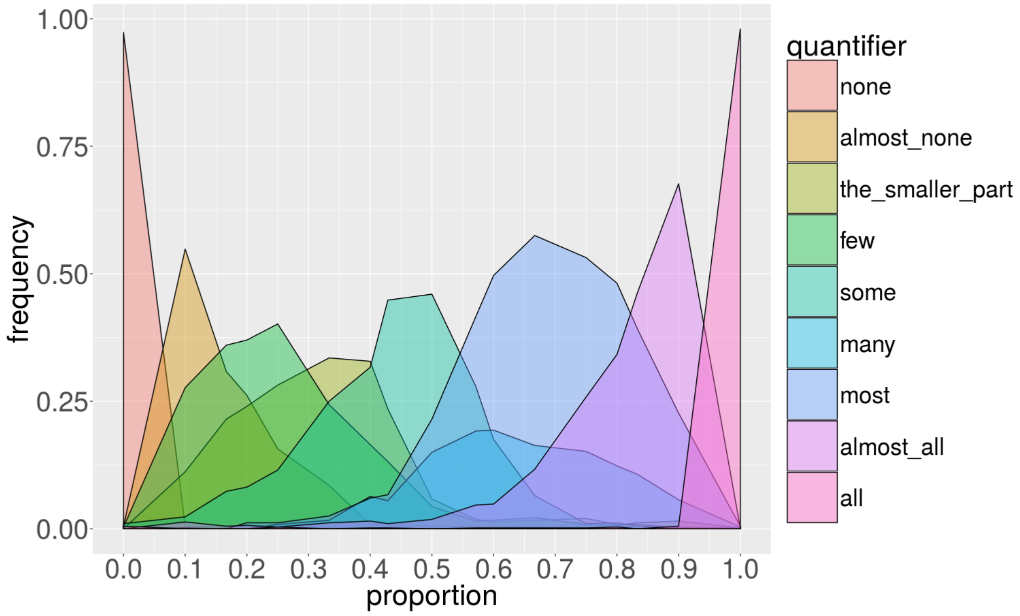

Figure 3: Density plot reporting the frequency distribution of responses for the 9 quantifiers (y-axis) against the proportion of targets in the scene (x-axis).

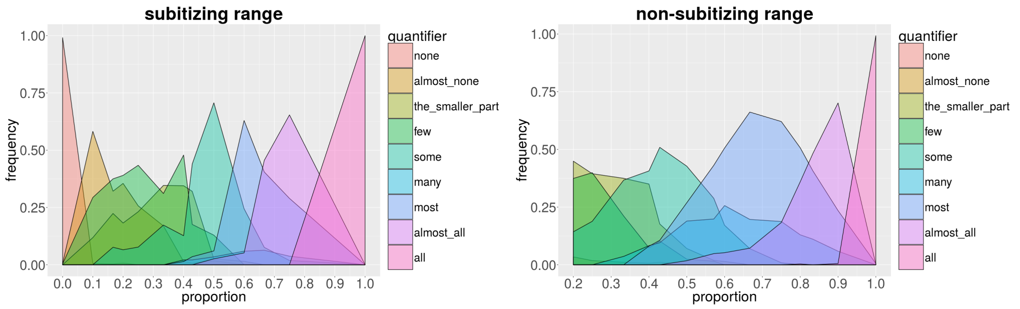

Figure 4: Density plots reporting frequency distribution of responses against proportion of targets for scenes whose number of targets is within the subitizing range (left) and exceeding it (right).

targets and cardinality of both targets and non-targets), it was not possible to disentangle between the relative contribution of the two (or more) factors within the same logistic regression model. We thus employed the "one model, one predictor" strategy, according to which a number of separate models including only one predictor of interest (along with random factors) was performed for each quantifier. This way, the predictive power of each variable could be tested separately, and we could further evaluate the quality of each model relative to all other candidate models. Model selection was performed using Akaike Information Criterion (AIC), a measure based on information theory which allowed us to select the best model for a given set of data (Akaike, 1973). In particular, the lowest the AIC, the lowest the information loss compared with the "true" model, namely the process that generated the data. We considered both raw AIC scores and AIC weights (Wagenmakers & Farrell, 2004).

Seven variables were used as predictors: (a) proportion of targets, (b) cardinality of targets, (c) cardinality of non-targets, (d) subitizing/non-subitizing range (dichotomic dummy variable), (e) average size of targets, (f) average size of non-targets[2]. In total, 52 models were tested. All models were mixed-effect logistic regressions (Baayen et al., 2008) with one fixed predictor (see above) and 3 random factorial variables, namely (1) participant, (2) experimental block, and (3) position of the quantifier in the response grid. By including these random variables in the models, we ensured that significant effects were estimated for the whole set and not just for a sample of stimuli. That is, we ensured that the effects were not due to the variability among participants, blocks of stimuli, position of the quantifier word in the response grid. To better fit the data, all the models except (d) treated the predictor as a second-order polynomial variable. Logit models were performed using the function `lmer()` implemented in the package `lme4`.

To compare different models, raw AIC scores and AIC weights were used. Since, in all cases, AIC weights for the lowest-AIC model approximated 1 (i.e. the total weight of the models considered), Table 2 reports only AIC scores for all models. As can be seen, for 8 quantifiers out of 9, the best model (i.e. the one with the lowest information

[2] The average size of the targets was obtained by dividing their weigthed sum (each large target was multiplied by 1, medium ones by 0.75, small ones by 0.5) by the number of targets in the scene. The same criteria and procedure were used for non-targets. For intuitive reasons, scenes containing either 0 targets or 0 non-targets were excluded from this analysis.

quantifier	**AIC scores**					
	(a) % targ	(b) n targ	(c) n non-targ	(d) sub/non-sub	(e) targ size	(f) non-targ size
none	**613.03**	756.31	3474.44	3113.78	–	3913.39
almost none	4353.51	**4230.42**	5591.74	4292.00	4686.18	5633.89
few	**5492.00**	6015.22	6486.62	6428.88	6987.15	7018.23
the smaller part	**5241.33**	5938.05	6109.82	6605.17	6540.34	6451.78
some	**5811.28**	6864.85	7342.64	7792.71	7608.18	7461.32
many	**4273.67**	4520.02	4834.66	4600.70	4909.47	5062.78
most	**6755.09**	8402.49	8741.28	8748.20	9330.23	9604.31
almost all	**5079.70**	6355.7	5692.78	6545.65	6762.34	6075.14
all	**482.37**	3323.29	732.50	3672.75	3568.47	–

Table 2: AIC scores for each of the models. **Bold** values (lowest) correspond to best models. Empty cells indicate cases for which the number of datapoints was too low to perform statistical analyses.

loss) turned out to be the one using proportion of targets (% targ). In one case, namely "almost none", the best model was instead the one using cardinality of targets (n targ) as the predictor. The models based on all other predictors (cardinality of non-targets, subitizing/non-subitizing range, and either targets or non-targets average size) never emerged as the best ones for any quantifier.

It is worth stressing that AIC scores do not say anything about the absolute quality of the model, i.e. the testing of the null hypothesis. Once established the best models based on the AIC score, we could inspect them using the traditional null-hypothesis testing. For all best models, both the linear and the quadratic term of the polynomial variable turned out to be highly significant ($p < .0001$), meaning that each quantifier can be reliably predicted against the other quantifiers by means of the polynomial form of the given predictor. In Table 3, we report Estimate, z-value and p-value of the quadratic term (2nd order term) for each of the selected models.

Based on the well-reported effects due to subitizing, we analyzed separately the datapoints within the subitizing range, i.e. cardinality of targets up to 3 included. The intuition behind that is that when the target items are very easily enumerable (in the subitizing range), their absolute number might be a better predictor of the quantifier used by subjects than the proportion. To test this hypothesis, the same kind of analysis as above was performed on the split data ($3,771$ datapoints). For all quantifiers except "almost all", the best models turned out to be the polynomial ones, whereas for "almost all" the best model was the linear one. Table 4 reports AIC score, Estimate, z-value, and p-value of the quadratic term (linear term for "almost all") for the best models in the subitizing range. As can be noticed, in the subitizing range the low-magnitude quantifiers “none”, “almost none”, and “few” are better modeled by the absolute number of animals rather than by the proportion of targets. This suggests that the choice of these quantifiers in this range relies more on evaluating the set of targets on its own than comparing it against the set of non-targets.

Finally, we investigated whether the frequency of use of quantifiers in language is reflected in the distribution of responses observed in the experiment. The rationale is that, when choosing a quantifier from the various options, participants might be biased towards the most frequent words, irrespectively of the perceptual features of the visual stimulus. We extracted raw frequency values for each of the 9 Italian quantifiers at the lemma level from CORIS (Favretti et al., 2002) and we computed the Pearson's correlation (r) with the quantifier frequencies observed in the experiment. All the values were previously log-transformed. The correlation turned out to be very weak and not significant in the full dataset ($r(7) = -0.25$, p=0.52), in the subitizing range subset ($r(7) = -0.41$, p=0.27), and in the non-subitizing range subset ($r(7) = -0.04$, p=0.92). That is, participants are not affected by the linguistic frequency of the quantifier when picking it up from the list.

quantifier	**predictor**	**Estimate**	**z-value**	**p-value**
none	proportion	424.78	19.36	.0001
almost none	n targets	82.86	9.66	.0001
few	proportion	-215.02	-22.41	.0001
the smaller part	proportion	-235.73	-25.98	.0001
some	proportion	-279.16	-35.69	.0001
many	proportion	-210.73	-6.31	.0001
most	proportion	-288.99	-29.79	.0001
almost all	proportion	-147.51	-13.67	.0001
all	proportion	462.95	18.66	.0001

Table 3: Estimate, z-value and p-value of the quadratic term for each of the best models.

quantifier	**predictor**	**AIC score**	**Estimate**	**z-value**	**p-value**
none	n targets	328.2	158.15	11.41	.0001
almost none	n targets	2572.3	-136.47	-20.35	.0001
few	n targets	3541.3	-69.75	-13.84	.0001
the smaller part	proportion	2662.3	-110.61	-13.40	.0001
some	proportion	2057.6	-88.07	-12.69	.0001
many	proportion	256.9	-195.17	-4.38	.0001
most	proportion	733.8	-57.04	-4.74	.0001
almost all	proportion	629.2	8.97	13.81	.0001
all	proportion	57.8	247.04	2.72	.0064

Table 4: AIC score, estimate, z-value and p-value of the quadratic term (linear term for "almost all") for each of the best models in the subitizing range.

3.2. Abstract task: semantic similarity judgements

The pattern of estimated similarities across quantifiers indicated that quantifiers are represented on an ordered but highly non-linear scale. A visualization of that can be observed in Figure 5, where a heatmap depicting the average semantic similarity between quantifier pairs is reported. Three interesting features can be appreciated: First, the ordered aspect of the internal scale can be seen by observing a roughly graded decrease in similarity as pairs move away from the diagonal. This indicates a rough "distance effect", indexing an internal ordered scale. This distance effect appears stronger for low-magnitude quantifiers compared to high-magnitude ones. This can be appreciated qualitatively by inspecting Figure 6, where the bell functions peaking around the low-magnitude quantifiers ("few", "the smaller part", "almost none", "none") appear sharper compared to those characterizing the high-magnitude quantifiers ("many", "most", "almost all", "all").

Second, it appears that this graded effect is mostly confined in quantifiers that refer to similar magnitudes, and disappears for very distant quantifiers. Indeed, there seems to be a clear-cut distinction between low-magnitude and high-magnitude quantifiers. In this respect, "some" turns out to be a "hinge" between low- and high-magnitude quantifiers. It should be observed that none of the items are judged to be as extremely similar/dissimilar to it, with the lowest average similarity being equal to 3.08 ("all-

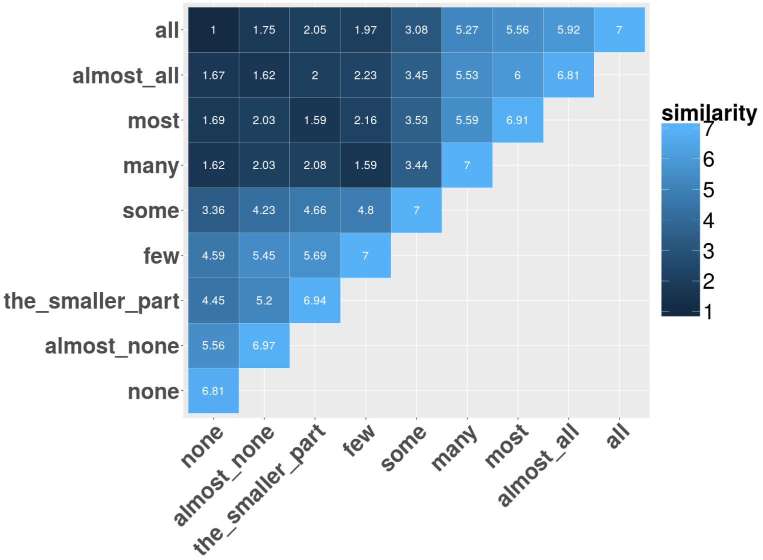

Figure 5: Heatmap reporting the average semantic similarity between quantifiers pairs. The lighter the blue, the more similar the pair.

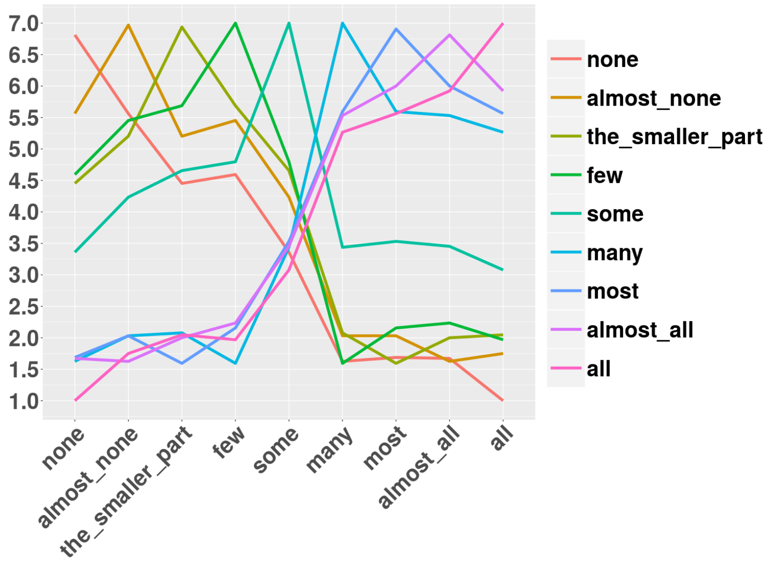

Figure 6: Line plot reporting the average semantic similarity between quantifiers.

some"), and the highest being equal to 4.8 ("few-some"). Though halfway between low- and high-magnitude quantifiers, however, "some" results to be closer to the former than to the latter group. Finally, we observe a rather small but systematic linguistic "antinomy effect": For each quantifier (with the exception of "some") the most dissimilar item is represented not by the extreme on the other side of the scale, but by its linguistic *antonym*: The lowest similarity ratings are those among "none-all", "almost none-almost all", "the smaller part-most", "few-many" (this can be appreciated by the presence of an orthogonal diagonal to the main one in the similarity matrix).

To pool together the pattern of judgements and reconstruct the shape of the internal representation, we performed a metric Multi-Dimensional Scaling (MDS) analysis. Figure 7 shows the results of the analysis when taking into account two dimensions. By performing a goodness-of-fit analysis, it turned out that the first dimension only, depicted along the x-axis in the plot, accounts for 98.66% of the variance of the original data (R^2=0.9866, *F(*1, 34*)*=2496.81, p< .0001). As shown in Figure 7, such dimension clearly separates low-magnitude quantifiers from high-magnitude quantifiers, with "some" somehow in between, though closer to the former block. By including the second dimension, the variance accounted for by the model increases to 98.80% (R^2=0.9880, *F(*1, 34*)*=2803.18, p< .0001), which is almost a perfect fit. Such dimension neatly represents magnitude: From low to high, along the y-axis. This analysis further confirms that low-magnitude quantifiers are better separated among them, indicating that they correspond to sharper representations. This allows their ordering on a scale to emerge very clearly, with "none" being followed by "almost none" that, in turn, is followed by "few" and "the smaller part" (which are not well separated among each other), and eventually by "some". On the contrary, high-magnitude quantifiers, while still being ordered along a scale, are extremely close to each other, indicating that their representations overlap greatly.

4. Discussion

4.1. Visually-grounded representation: Proportion, cardinality and object size

In this paper, we explored the use of quantifiers in both their visually-grounded and abstract representation. By asking participants to choose the quantifier that best

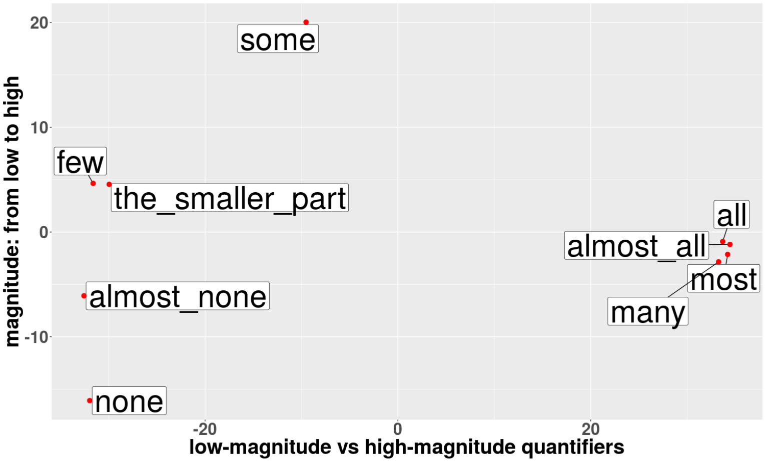

Figure 7: Plot reporting the absolute distance of quantifiers as resulting from a two-dimension metric MDS analysis.

represented the quantity of animals in a number of visual scenes, Experiment 1 was aimed at investigating the factors which contribute in determining the visually-grounded representation of such linguistic expressions. We showed that the proportion of targets is the best predictor for 8 quantifiers out of 9, with “almost none” being better described by the cardinality of the target set. When zooming into the subitizing range, with cardinality of animals up to 3, the absolute number of targets turned out to be the best predictor for “none” and “few” besides “almost none”, thus suggesting that when the information about precise number is available it becomes crucial for discriminating among low-magnitude quantifiers.

These findings are generally in line with previous studies investigating the appropriateness of quantifiers evaluated against visual scenes (Coventry et al., 2005, 2010). Using a different experimental design (evaluating the appropriateness of a number of quantifier-embedding sentences against a given visual scene), a different set of quantifiers (“a few”, “few”, “several”, “many”, “lots of”), and without constraining the exposure time to the scene, these works showed that the number of both targets and non-targets is predictive of the quantifier appropriateness. With cardinality of targets equal to 3 (their subitizing case), however, the use of quantifiers was no longer affected by the cardinality of the non-target objects. An exception was represented by “few”, which was affected by both

(Coventry et al., 2010). On the one hand, our finding that proportion is overall the best predictor is not in contradiction with the effect of both number of targets and number of non-targets. Rather, we believe ours to be just a better measure to assess the contribution of both sets in determining quantifiers' use. On the other hand, the results we obtained in the subitizing range reinforce and better prove the increasingly important role of precise number in discriminating between low-magnitude quantifiers. In our study, interestingly, the only low-magnitude quantifier whose interpretation turned out to be best predicted by the proportion of targets also in the subitizing range was "the smaller part", whose reading is intuitively more proportional compared to the others. Finally, it is worth stressing that our 340 visual scenes were balanced with respect to ratios, whereas the 36 used by Coventry et al. (2005, 2010) were balanced for target cardinality. Moreover, in the present work each ratio was represented by all possible combinations of cardinalities, whereas Coventry and colleagues experimented with ratios that were mostly depicted by just one combination. Finally, our subitizing range included four cardinalities, namely 0, 1, 2, and 3 – not just the number 3.

As far as the effect of object size is concerned, we found this factor not to be among the most predictive ones. This finding is in partial contrast with the results reported by Newstead & Coventry (2000), who showed a role of size in the task of evaluating quantifiers over scenes depicting dots placed in a container. In that study, both the dots and the container size were found to play a role: Low-magnitude quantifiers were found to be more appropriate when the dots were small and when the container was big. In our task, we solely investigated the size of the items, and found that this parameter was not among the best predictors of quantifiers' use. This difference might be due to the different experimental settings: First, our scenes contain both target and non-target objects – not only targets. Second, we vary the size of the objects in a way that there are no scenes depicting, e.g., only small or large objects. Third, we employ a larger set of quantifiers, thus participants have more alternatives compared to the previous study. Moreover, contrary to us, Newstead & Coventry (2000) allowed subjects to explore the scenes for an infinite time, such that they might have used a different visuo-spatial strategy (namely, exact counting), and that might have influenced the enumeration process.

4.2. Abstract representation

By asking participants to rate the degree of semantic similarity between quantifier pairs, Experiment 2 was aimed at testing whether these expressions are mentally ordered and, if so, which are the features of the resulting scale. We showed that, even without relying on any quantitative or contextual information, quantifiers do lie on an ordered scale, as resulting from a Multi-Dimensional Scaling Analysis (Kruskal & Wish, 1978). In particular, low-magnitude quantifiers ("none", "almost none", "few", "the smaller part") turned out to be perceived as being fairly distant from each other, thus suggesting that their abstract semantic representation is well defined and nicely ordered on a scale. In contrast, high-magnitude quantifiers ("many", "most", "almost all", "all") turned out to greatly overlap, though always along an ordered scale. Overall, these results suggest that the mental representation of quantifiers is ordered and highly non-linear, with small quantifiers better represented compared to large ones. This is highly reminiscent to the well-reported logarithmic scale inferred both from comparative judgements across numerical symbols and from the use of numerical symbols in perceptual quantification (Nieder & Miller, 2003; Dehaene, 2003; Dehaene et al., 2008).

It is worth stressing that, in doing this task, neither quantitative (numbers, proportions, etc.) nor explicit contextual (semantic) information was provided. That is, quantifiers were judged in isolation, solely on the basis of their bare semantic similarity, while in Holyoak & Glass (1978) participants were asked to rate dissimilarities between pair of sentences embedding different quantifiers. Another interesting finding was the tendency to assign the lowest rating (i.e. lowest semantic similarity) to the direct antonym. For example, the most dissimilar word from "few" was "many", and not "none". While straightforward for the pair "none-all", which also represent the two extreme endpoints of the scale, this finding is in principle not trivial in all the other cases. This finding falls off the prediction that quantifiers should solely lie on a quantitative scale (e.g. numerical or proportional) and suggests that, when asked to judge the semantic similarity of a word pair, speakers also take into account lexico-semantic features, such as information regarding the direct antonym (Miller & Fellbaum, 1991), as also reported by Hill et al. (2016).

4.3. Mental order

Finally, it should be mentioned that previous work has investigated the scalar nature of quantifiers from very different perspectives. With a set of 5 quantifiers and a task which was similar to ours, for example, Holyoak & Glass (1978) claimed that quantifiers can be described in terms of an unidimensional scale, essentially representing analog quantities. The authors, however, did not overtly exclude that information regarding other non-quantitative related semantic features might still be included in the memory representation of quantifiers. In contrast with the unidimensional nature of the quantifier scale was Routh (1994), whose results on a freesort task with 20 quantifiers suggested that several other components are in place beyond the quantity scale. Another study (Montalto et al., 2010) also adopted a similar paradigm where a number of Italian quantifiers (yet different from the list of quantifiers investigated in our study) were compared to each other on a magnitude scale: Given pairs of quantifiers subjects had to indicate if and which of the two indicated the largest amount. Differently from our experiment, however, subjects were given the possibility to indicate that the two quantifiers did not differ in the implicated amount. Results suggested that subjects lump quantifiers in two blocks, one comprising low and the other high-magnitude ones, with no hint of a continuous scale. However, there is the serious possibility that these results do not directly reflect the true mental scale but rather the degree of certainty, such that when prompted with uncertain decisions subjects indicated an absence of differentiation.

4.4. Impact of our results on foundational theories

As for the theoretical implications of our work, our results provide evidence in support of some well-established assumptions on quantifiers. First, our findings show that quantifiers neither correspond to an exact number of entities nor to a fixed proportion. This can be taken as an evidence in favor of their non-referential status, even in the new light shed by the integration of perception and quantifiers.

Second, our results do not shed new light on the proposal that “few” and “many” are not-extensional since, in our experiments, contextual factors were deliberately avoided. However, it is worth noticing that in Experiment 1 the meaning of “few” is found to be ambiguous: It mostly depends on the number of targets in the subitizing range, on the proportion of targets in the whole data. This might be seen as an effect of a perceptual "contextual" factor: "Few" is more dependent on the perceptual context than are other quantifiers. However, the same effect was not observed for "many".

Third, our results are consistent with the literature on scalar implicatures (Grice, 1975) in the pragmatic use of quantifiers. In particular, both the ordering of quantifiers (from low- to high- magnitudes) and their narrow range of application observed in Experiment 1 suggest that, to some extent, speakers interpret such expressions as having an upper boundary which excludes the use of more informative options when these options are not true or uncertain (Horn, 1984). That is, participants choose the most informative quantifier "all" (and not e.g. "some", which would be logically true) when they are certain about its applicability. Similar implications can be drawn from Experiment 2, where the characteristics of the abstract representation might indicate that speakers have an internal representation of quantifier informative strength. Based on our findings, one possibility is that scalar implicatures are stronger for low-magnitude quantifiers (which turn out to be extremely well-defined and distinct from each other) than for high-magnitude ones (which are perceived to be very similar). We leave this issue for future research and refer the reader to Oaksford et al. (2002) for interesting results on the use of quantifiers as referring to different ranges of numerosities and their effect on informativeness.

Fourth, the results of Experiment 2 are in line with the position that the meaning of quantifiers is not only about amounts, numbers, or proportions. Indeed, similarity judgments provided by participants turned out to be dependent on lexico-semantic factors (e.g. antonymy) besides magnitude. This evidence is also in line with previous findings showing an interplay between numerical and semantic information in the comprehension of quantifiers (Heim et al., 2012).

Fifth, our results overall suggest that the meanings of quantifiers are at least partially tied to the representation of quantities. Though this is probably not enough to devise a general semantics for such expressions, we believe quantitative aspects to constitute the basis of quantifier meanings.

4.5. Final remarks

In sum, our results indicate that quantifiers primarily represent proportions and not absolute cardinalities, apart from when they refer to sets with less than four objects. They also show that quantifiers are mentally represented on a quantity scale which is well ordered and highly non-linear, bearing interesting similarities to the mental representation of both numerical quantities and continuous magnitudes. While our results cannot endorse one possibility over the other, they firmly support the view that quantifiers are mentally represented in a way that partially reflects the way we perceive quantities through our senses.

Acknowledgements

We kindly acknowledge Marco Marelli and Roberto Zamparelli for their valuable insights and feedback. We are also grateful to Marta Cristofanini for her contribution in collecting data of Experiment 2. The first Author gratefully acknowledges the iV&L Net (ICT COST Action IC1307) for funding his short-term research visit, which allowed him to get useful feedback and comments.

References

Akaike, H. (1973). Information Theory and an extension of the Maximum Likelihood Principle. In *2nd International Symposium on Information Theory, Budapest: Akademiai Kaido* (pp. 267–281).

Arnold, J. B. (1971). A multidimensional scaling study of semantic distance. *Journal of Experimental Psychology*, *90*, 349.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, *59*, 390–412.

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and philosophy*, *4*, 159–219.

Coventry, K. R., Cangelosi, A., Newstead, S., Bacon, A., & Rajapakse, R. (2005). Grounding natural language quantifiers in visual attention. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society. Mahwah, NJ: Lawrence Erlbaum Associates*.

Coventry, K. R., Cangelosi, A., Newstead, S. E., & Bugmann, D. (2010). Talking about quantities in space: Vague quantifiers, context and similarity. *Language and Cognition*, *2*, 221–241.

Dehaene, S. (2003). The neural basis of the Weber–Fechner law: A logarithmic mental number line. *Trends in cognitive sciences*, *7*, 145–147.

Dehaene, S., Izard, V., Spelke, E., & Pica, P. (2008). Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures. *Science*, *320*, 1217–1220.

Deschamps, I., Agmon, G., Loewenstein, Y., & Grodzinsky, Y. (2015). The processing of polar quantifiers, and numerosity perception. *Cognition*, *143*, 115–128.

Favretti, R. R., Tamburini, F., & De Santis, C. (2002). CORIS/CODIS: A corpus of written italian based on a defined and a dynamic model. *A rainbow of corpora: Corpus linguistics and the languages of the world*, (pp. 27–38).

Graves, R., & Hodge, A. (1943). The reader over your shoulder.

Grice, H. P. (1975). Logic and conversation – in Syntax and Semantics. Vol. 3, Speech Acts. *ed. Peter Cole and Jerry Morgan*, (pp. 41–58).

Hammerton, M. (1976). How much is a large part? *Applied ergonomics*, *7*, 10–12.

Heim, S., Amunts, K., Drai, D., Eickhoff, S., Hautvast, S., & Grodzinsky, Y. (2012). The language–number interface in the brain: a complex parametric study of quantifiers and quantities. *Frontiers in evolutionary Neuroscience*, *4*, 4.

Hill, F., Reichart, R., & Korhonen, A. (2016). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, .

Holyoak, K. J., & Glass, A. L. (1978). Recognition confusions among quantifiers. *Journal of verbal learning and verbal behavior*, *17*, 249–264.

Horn, L. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. *Meaning, form, and use in context: Linguistic applications*, (pp. 11–42).

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

Keenan, E. L., & Stavi, J. (1986). A semantic characterization of natural language determiners. *Linguistics and philosophy*, *9*, 253–326.

Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of neurophysiology*, *97*, 4296–4309.

Kruskal, J., & Wish, M. (1978). Quantitative applications in the social sciences: Multidimensional scaling (vol. 11). Beverly Hills.

Miller, G. A., & Fellbaum, C. (1991). Semantic networks of English. *Cognition*, *41*, 197–229.

Montague, R. (1973). The proper treatment of quantification in ordinary English. In *Philosophy, language, and artificial intelligence* (pp. 141–162). Springer.

Montalto, R., Van Hout, A., & Hendriks, P. (2010). Comparing children's and adults' interpretation of Italian indefinite quantifiers. *Linguistics in Amsterdam*, *3*, 1–19.

Moxey, L. M., & Sanford, A. J. (1993). Prior expectation and the interpretation of natural language quantifiers. *European Journal of Cognitive Psychology*, *5*, 73–91.

Moxey, L. M., & Sanford, A. J. (2000). Communicating quantities: A review of psycholinguistic evidence of how expressions determine perspectives. *Applied Cognitive Psychology*, *14*, 237–255.

Newstead, S., Pollard, P., & Riezebos, D. (1987). The effect of set size on the interpretation of quantifiers used in rating scales. *Applied Ergonomics*, *18*, 178–182.

Newstead, S. E., & Coventry, K. R. (2000). The role of context and functionality in the interpretation of quantifiers. *European Journal of Cognitive Psychology*, *12*, 243–259.

Nieder, A., & Miller, E. K. (2003). Coding of cognitive magnitude: Compressed scaling of numerical information in the primate prefrontal cortex. *Neuron*, *37*, 149–157.

Nouwen, R. (2010). What's in a quantifier? *The Linguistics Enterprise: From knowledge of language to knowledge in linguistics*, *150*, 235.

Oaksford, M., Roberts, L., & Chater, N. (2002). Relative informativeness of quantifiers used in syllogistic reasoning. *Memory & cognition*, *30*, 138–149.

Paterson, K. B., Filik, R., & Moxey, L. M. (2009). Quantifiers and discourse processing. *Language and Linguistics Compass*, *3*, 1390–1402.

Piazza, M., & Eger, E. (2016). Neural foundations and functional specificity of number representations. *Neuropsychologia*, *83*, 257–273.

Routh, D. A. (1994). On representations of quantifiers. *Journal of Semantics*, *11*, 199–214.

Schöller, A., & Franke, M. (2017). Semantic values as latent parameters: Testing a fixed threshold hypothesis for cardinal readings of few & many. *Linguistics Vanguard*, *3*.

Shikhare, S., Heim, S., Klein, E., Huber, S., & Willmes, K. (2015). Processing of numerical and proportional quantifiers. *Cognitive science*, *39*, 1504–1536.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2004). Word association spaces for predicting semantic similarity effects in episodic memory. *Experimental cognitive psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer*, (pp. 237–249).

Szabolcsi, A. (2010). *Quantification*. Cambridge University Press.

Van Benthem, J. et al. (1986). *Essays in logical semantics*. Springer.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic bulletin & review*, *11*, 192–196.

Westerståhl, D. (1985). Determiners and context sets. *Generalized quantifiers in natural language*, *1*, 45–71.

Yildirim, I., Degen, J., Tanenhaus, M., & Jaeger, F. (2013). Linguistic variability and adaptation in quantifier meanings. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 35.