Truth and typicality in the interpretation of quantifiers[*]

Bob VAN TIEL — *Universität Bielefeld*
Bart GEURTS — *Radboud University Nijmegen*

Abstract. The standard view in natural language semantics is that quantifiers denote relations between sets. Psychological research, however, has shown that quantified statements often convey finer-grained information that is not encoded in their set-theoretic definitions. We investigate the relationship between these meaning aspects on the basis of two experiments.

Keywords: quantification; probability; typicality; scalar inference.

1. The interpretation of quantifiers

The interpretation of quantifiers has been investigated from a range of perspectives. The standard view in natural language semantics is that quantifiers denote relations between sets (e.g., Barwise and Cooper 1981; Keenan and Stavi 1986; Montague 1973). For example, 'All A are B' is true iff the set of A is a subset of the set of B, 'Some A are B' is true iff the intersection between the set of A and the set of B is nonempty, and 'Most A are B' is true iff the number of A that are B is greater than the number of A that are not B. Formally, using the italic form of a term to refer to its extension:

$$\begin{array}{rll} \text{'All A are B'} & \text{is true iff} & A \subseteq B \\ \text{'Some A are B'} & \text{is true iff} & A \cap B \neq \emptyset \\ \text{'Most A are B'} & \text{is true iff} & |A \cap B| > |A - B| \end{array}$$

These set-theoretic definitions assign binary truth values: quantified sentences are either true or false in a situation. No finer-grained differences between situations are thus expected.

Psychological research, however, suggests that quantified statements often convey finer-grained information than what is encoded in their set-theoretic definitions. To illustrate, Newstead et al. (1987) asked participants to fill in the blanks in sentences like the following, where the quantifier Q and the total set size *n* were varied between items:

If Q of a group of *n* people are male, then _____ people are male.

In addition, Newstead et al. asked participants what they would expect to be the minimum and maximum number of people that satisfied the predicate given the truth of the antecedent. For

[*]This research was supported by a grant from the Netherlands Organization for Scientific Research (NWO), which is gratefully acknowledged.

statements where the total set size was 60, Newstead et al. found the following mean estimates (in %) for the previously mentioned quantifiers ‘all’, ‘some’, and ‘most’:

	Min	*Mean*	*Max*
All	100	100	100
Some	17	33	45
Most	66	83	90

The results for ‘all’ are in line with its set-theoretic definition: ‘All people are male’ implies that there are no females. The response ranges for ‘some’ and ‘most’, however, are much smaller than suggested by their set-theoretic definitions. For example, even though, according to its set-theoretic definition, ‘Some people are male’ is true whenever more than one person is male, participants infer from this statement that between 17% and 45% people are male. Further research has shown that the precise estimates are influenced by extralinguistic factors, such as:

- *Total set size*. Newstead et al. (1987) found for some quantifiers that the estimated number of people that satisfies the predicate depends on the total set size. An example is ‘Some people are male’: if there are just twelve people, participants estimate that 37% of them are male, whereas in a situation with ten thousand people their estimation drops to about 27%.
- *World knowledge*. Estimates for statements of the form ‘Q A are B’ depend on the intuitive likelihood that As are B: they will be higher for statements like ‘Q people find Miss Sweden attractive’ than for statements like ‘Q earthquakes occurred in California in 1951’ (e.g., Moxey and Sanford 1993; Pepper and Prytulak 1974).
- *Audience design*. Yildirim et al. (2013) provide evidence that listeners tailor their interpretation of quantified statements to the idiosyncracies of the speaker. If a speaker consistently refers to situations where half of the A are B with ‘Some A are B’ instead of ‘Many A are B’, listeners take this information into account in their estimates.
- *Alternatives*. Chase (1969) found that estimates for quantified statements depend on the alternative expressions that feature in the experiment. He asked participants to rate the likelihood of an event on a five-point scale. In one condition, these events were described by means of high-frequency quantifiers (e.g., ‘very often’, ‘usually’); in the other condition, by means of low-frequency quantifiers (e.g., ‘seldom’, ‘occasionally’). In many cases, Chase found that the mean likelihood ratings in these conditions were statistically indistinguishable.

These findings can be modelled in various ways. Some authors have proposed that quantifiers denote probability distributions over situations. In other words, ‘Q A are B’ denotes a function from situations to numerical values that sum to one (e.g., Yildirim et al., 2013). However, there are a number of issues with this proposal. First, it is not immediately obvious what the numerical values represent. Suppose that the function assigns a value *p* to a particular situation. One interpretation is that this means that the listener believes that the likelihood of this situation is *p*. Another interpretation is that the likelihood that a listener believes this situation is the one the speaker had in mind is *p*. Yet another interpretation is that the listener believes that the speaker believes that the likelihood of that situation is *p*. For our current purposes, the differences between these proposals are immaterial but it is an issue that stands in need of further analysis. A more pressing problem with the probabilistic account is that for most quantifiers it presupposes knowledge about the size of the quantifier domain. Consider the sentence 'Some A are B'. In order to assign a probability value to a situation with, for example, five As that are B, it will be necessary to know how many As there are in total: the value will be much higher in a situation with ten As than in a situation with ten thousand As. This indeterminacy goes against the intuition that it is perfectly possible to interpret quantified statements without having knowledge about the size of the quantifier domain.

In order to avoid this issue, we will model the finer-grained interpretation of quantified statements by assigning them functions from situations to typicality values (Rosch, 1975). This is a more general approach than the probabilistic account because it does not require that the numerical values sum to one. It is therefore possible to assign a definite value to a situation even it the total set size is unknown. The numerical values represent the "typicality" of a situation with respect to the quantified statement. In the next section, we explain the notion of typicality in some more detail. Typicality values can be converted to probability values by dividing them by the sum of the typicality values, which will only be possible if information about the total set size is available.

What is the relationship between typicality structure and the set-theoretic truth definitions proposed by natural language semanticists? Are these meaning aspects disparate, or are they reflections of one underlying dimension? Are set-theoretic definitions of quantifiers still needed in light of the findings from psychological experiments? In Section 4, we address these questions on the basis of the results of two experiments that will be discussed in Section 3. This investigation follows the lead of McCloskey and Glucksberg (1978), who inquired into the interpretation of nouns. In the next section, we consider their arguments in some detail.

2. McCloskey and Glucksberg (1978)

Nouns like 'bird' and 'furniture' refer to categories. According to the classical view, already propagated by Socrates in Plato's *Statesman* and later popularised by Aristotle, categories are sets of objects that fulfill a list of necessary and sufficient conditions. For example, 'bird' refers to the set of individuals that are warm-blooded and egg-laying vertebrates with feathers and wings. According to this account, all individuals are either birds or nonbirds.

Psychological research, however, suggests that listeners often make finer-grained distinctions between objects than the binary distinction imposed by the classical definition of a category. For example, Rosch (1975) found that participants consider sparrows to be more typical birds than penguins or chickens. Some authors have argued that these typicality judgements indicate that category membership itself is a matter of degree: sparrows are birds to a greater degree than are penguins or chickens (e.g., Lakoff 1973). Other authors, however, have criticised this view (e.g., Kamp and Partee 1995).

What is the relationship between typicality judgements and category membership? Do listeners have access to classical definitions for categories denoted by nouns like ‘bird’ and ‘furniture’? To address these questions, McCloskey and Glucksberg (1978) probed participants for typicality judgements and category membership judgements for a range of categories and objects.

Table 1 provides a sample of the results for the categories denoted by ‘bird’ and ‘furniture’. In both cases, the average typicality values line up along a continuum between the two extremes. In the case of ‘furniture’, the percentages of positive responses in the category membership task also form such a continuum. In the case of ‘bird’, by contrast, almost all percentages of positive responses are close to the extremes of 0 and 100. Hence, there was more agreement about category membership for ‘bird’ than for ‘furniture’.

Object	$\boldsymbol{\tau}$	$\boldsymbol{\varsigma}$	***Object***	$\boldsymbol{\tau}$	$\boldsymbol{\varsigma}$
Bird			*Furniture*		
Robin	10.00	100	Chair	9.95	100
Eagle	9.58	100	Table	9.83	100
Partridge	8.42	100	Bed	9.58	98
Goose	8.29	97	Rug	6.25	48
Condor	8.23	100	Lampshade	5.70	63
Buzzard	8.08	98	Sewing machine	5.32	11
Turkey	7.92	100	Refridgerator	5.07	18
Chicken	7.75	95	Waste basket	4.70	31
Loon	7.43	100	Bookends	4.53	43
Ostrich	7.25	97	Ironing board	4.32	16
Penguin	6.96	92	Pillow	4.12	31
Bat	3.63	17	Electric fan	3.78	13
Flying squirrel	2.63	5	Ashtray	3.45	21
Vampire	2.29	13	Door	2.87	10
Bee	2.04	3	Ceiling	2.03	0
Locust	1.83	9	Fence	1.87	0

Table 1: Sample of the results of the typicality and category membership tasks for the categories denoted by ‘bird’ (left column) and ‘furniture’ (right column) as found by McCloskey and Glucksberg (1978). $\boldsymbol{\tau}$: Average typicality on a 10-point scale; $\boldsymbol{\varsigma}$: Percentage of positive responses in the category membership task.

Do participants have access to a well-defined category for these nouns, or are their category membership judgements fully determined by typicality differences? In order to answer this question, we constructed two models predicting proportions of positive responses in the category membership task: one based on a classical definition and one based on typicality judgements. The classical model for ‘bird’ assigned 1 to all biological birds and 0 to all other objects. In the case of ‘furniture’, there was no straightforward criterion for distinguishing category members from nonmembers. Therefore a cutoff point in the typicality ratings was used: all objects that scored higher than 5.5 were assigned 1 and all other objects 0. The typicality model was formed by the normalised typicality ratings. For both nouns, we compared the absolute differences between the predicted and attested proportions of positive responses in the category membership task by means of Welch *t*-tests. In the case of ‘bird’, the classical model provided a better fit than the typicality model (mean differences of .06 and .17, $t(54) = -3.94$, $p < .001$), whereas the converse was the case for ‘furniture’ (mean differences of .19 and .09, $t(34) = 2.48$, $p = .02$).

The results for ‘bird’ are thus in accordance with the classical account of categorisation. For this noun, judgements of category membership were relatively crisp. This suggest that participants have access to a well-defined category of birds. The results for ‘furniture’ are in accordance with the typicality account, since judgements of category membership were better approximated by typicality judgements than by any classical definition. These observations can be formalised as follows. Here, $\varsigma_{\text{A}}(x)$ is the proportion of participants who indicate that x is an instance of the category denoted by ‘A’, and $\tau_{\text{A}}(x)$ is the normalised mean typicality rating for x in the category denoted by ‘A’. These are values in the interval [0, 1]. $x \in \text{A}$ means that x is a member of the category denoted by ‘A’ according to its classical definition. This equals a value in the set $\{0, 1\}$.

$$\begin{aligned} \varsigma_{\text{BIRD}}(x) &= x \in \text{BIRD} \\ \varsigma_{\text{FURNITURE}}(x) &= \tau_{\text{FURNITURE}}(x) \end{aligned}$$

A further question that stands in need of an explanation is what determines the typicality judgements that McCloskey and Glucksberg found. In the case of ‘bird’, category membership plays a prominent role in the typicality judgements as well: the difference in mean typicality rating between the least typical birds (i.e., 6.96 for penguins) and the most typical nonbirds (i.e., 4.96 for pterodactyls) is much greater than the difference between any other pair of neighbours on the ς-scale. No such effect is visible in the case of ‘furniture’. In addition, typicality judgements are often explained in terms of distance from the prototype (e.g., Rosch and Mervis 1975). A prototype is an object that is especially representative of a category because it satisfies most or all of the characteristics that are standardly associated with that category. For example, a prototypical bird might be an animal that is capable of flight, relatively small, and not too exotic. These observations can be formalised as follows. Here, $dist(x, p)$ is a measure of the distance between x and the prototype p. This equals a value in the interval [0, 1]. The resultant typicality values occur in the interval [-1, 1] and should therefore be normalised to the [0, 1] interval.

$$\begin{aligned}\tau_{\text{BIRD}}(x) &= x \in \text{BIRD} - dist(x,p)\\ \tau_{\text{FURNITURE}}(x) &= dist(x,p)\end{aligned}$$

In order to address the questions we posed at the end of the Introduction, we conducted two experiments analogous to McCloskey and Glucksberg's to determine and model the relationship between set-theoretic definitions and typicality structure in the interpretation of quantifiers. To that end, we gathered and analysed truth value judgements and typicality judgements for quantified statements. Since sentences refer to situations instead of individuals, we used pictures of situations instead of words referring to individuals. An example of a trial is shown in Figure 1. The quantifiers that were included in the experiments are listed in Table 2.

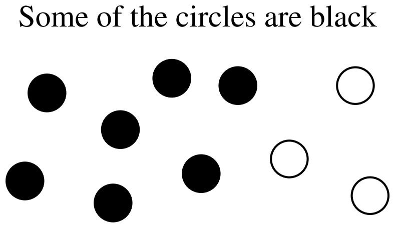


Figure 1: Sample item used in the experiments.

In the first experiment, participants had to indicate on a seven-point scale how well the situation was described by the statement; in the second experiment, they had to indicate whether the statement was true or false in the depicted situation. One of our goals was to investigate if truth value judgements are better approximated by set-theoretic definitions or by typicality judgements. The set-theoretic definitions we used to this end are also listed in Table 2.

Quantifier	***Definition***	***Quantifier***	***Definition***
All	$A \subseteq B$	Most	$\lvert A\cap B\rvert > \lvert A-B\rvert$
Every	$A \subseteq B$	None	$A\cap B = \emptyset$
Few	$\lvert A\cap B\rvert < \delta$	Some	$A\cap B \neq \emptyset$
Many	$\lvert A\cap B\rvert > \delta$	Not all	$A \not\subseteq B$
More than half	$\lvert A\cap B\rvert > \lvert A-B\rvert$	Not many	$\lvert A\cap B\rvert \leq \delta$

Table 2: Quantifiers used in the experiments and the corresponding set-theoretic definitions.

These definitions are the standard ones from the literature. 'Few' and 'many' are vague quantifiers. This implies that their meaning depends on the context. 'Few A are B' expresses that the number of A that are B is surprisingly low. This can be formalised by means of a contextually determined threshold value δ below which the real number of A that are B is supposed to fall. 'Many A are B' conversely implies that the number of A that are B is surprisingly high, and should therefore exceed some contextually determined threshold value δ. It turns out that in our experiment, the value that most participants assigned to δ was 5 for both quantifiers.

There is one likely complication that warrants some further discussion. It is well known that truth value judgements are influenced by pragmatic inferences. To illustrate, the truth conditions of 'Some A are B' are compatible with all situations where one or more A is B. Nonetheless, someone who utters this sentence will often exclude some of these situations by pragmatic means. This particular utterance carries at least three possible inferences. First, it might implicate that it is not the case that only one A is B. This inference is triggered by the plural marking on the subject and verb. Some authors have argued that this plurality inference is pragmatic in nature (e.g., Spector 2006). Second, the utterance might implicate that it is not the case that all of the A are B. This is a scalar inference based on the lexical scale ⟨some, all⟩. By using a weaker expression on this scale, the speaker implicates that she believes that using the stronger expression would have caused the utterance to be false. Third, it might implicate that it is not the case that most of the A are B. This scalar inference is based on the lexical scale ⟨some, most⟩. Zevakhina (2012) provides evidence that the 'not all' inference is more robust than the 'not most' inference.

It has been shown that some participants judge sentences false in situations where their pragmatic inferences are false (e.g., Bott and Noveck 2004). Note that this does not necessarily mean that these participants consider a sentence like 'Some A are B' equally bad in a situation where all of the A are B as they do in a situation where none of them are, and where the set-theoretic truth conditions of the sentence are thus violated: Katsos and Bishop (2010) show that most participants consider the sentence worse in the last-mentioned situation when given the option to distinguish between degrees of badness. In our experiments, these differences in degrees of badness might reveal themselves in the typicality judgements.

Most of the quantifiers in our investigation licensed pragmatic inferences. The quantifiers 'many', 'more than half', and 'most' license the inference that not all of the circles are black; the quantifiers 'few', 'not all', and 'not many' license the inference that at least one of the circles is black. Note that 'some' is exceptional in that it carries three potential pragmatic inferences, whereas all of the other quantifiers have just one possible inference. It seems plausible to suppose that these pragmatic inferences will have an effect on the results of both experiments. We will discuss this issue in more detail in the Results section.

3. The experiments

3.1. Participants

We posted surveys for 340 participants on Amazon's Mechanical Turk. Only workers with an IP address in the United States were eligible for participation. In addition, these workers were asked to indicate their native language. Payment was not contingent on their response to this question. 120 participants provided truth value judgements (mean age: 34; range: 18-61; 68 females). All of these participants were native speakers of English. 220 participants provided typicality judgements (mean age: 37; range: 18-70; 135 females). 30 participants provided typicality judgements for the quantifiers 'some' and 'every'; 20 participants provided typicality judgements for all of the other quantifiers. 5 participants provided typicality judgements for more than one quantifier. 11 of the 220 participants in the typicality task were excluded from the analysis because they were not native speakers of English.

3.2. Materials

Sentences were of the following form:

(1) Q {circle is / of these circles are} black.

Q was instantiated by the quantifiers in Table 2. The corresponding pictures consisted of ten circles which were either black or white. The distribution of black and white circles was manipulated, thus creating eleven situation $s_0, \ldots, s_{10}$. In a situation s_n, n of the circles were black and the remaining $10-n$ circles were white. An example trial is shown in Figure 1.

Surveys in the truth value judgement task consisted of twenty trials. Each of the ten quantifiers was instantiated twice with two different pictures. The pictures for one quantifier always differed in at least three circles. Because there was an uneven number of situations, one of them, s_5, occurred twice as often as the other situations: 40 instead of 20 times. The order of the items was randomised for each participant, making sure that the same quantifier never occurred consecutively. The truth value judgement task differed from McCloskey and Glucksberg's who asked participants to categorise all possible instantiations. We avoided this procedure because it might lead to contrastive readings of quantifiers. Intuitively, if 'some' receives a contrastive reading, which usually manifests itself by means of prosodic stress, it excludes 'all' by entailment. We wanted to avoid this potential confound as much as possible.

In the typicality experiment, participants were presented with one quantifier in all eleven situations. The order of the items was randomised for each participant.

3.3. Procedure

To collect truth value judgements, participants were presented with the following instructions:

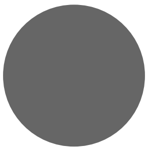
> In the following survey, we will show you pairs of pictures and sentences. In each case, we ask you to decide whether or not the sentence gives a correct description of the picture. If you feel that the sentence is true, check "True". If not, check "False".

> We are interested in your spontaneous judgments, so please don’t think too long about your answers.

In order to collect typicality judgements, participants were presented with the following instructions (based on the instructions used by Rosch 1975):

> This experiment is about how sentences are interpreted. Consider the sentence “This is a vehicle”. Many people agree that this sentence is a better description of a car or a motorbike than of a sled or a tractor, even though they are strictly speaking all vehicles. Below is another example.

This circle is dark.



bad ○ ○ ○ ○ ● ○ ○ good
1 2 3 4 5 6 7

> In my eyes, the picture is a reasonable instance of the sentence. I can imagine worse instances (for example a white circle) but I can also imagine better instances (for example a black circle). For that reason, I gave a rating that is in between the two extremes 1 and 7. However, the exact rating is a matter of taste and you might want to give a higher or lower rating. In this experiment, you will see one sentence with eleven pictures. For each picture, you have to indicate how well it is described by the sentence. It doesn’t matter why you think that a sentence is a good or bad description of a particular picture. Just follow your intuition.

3.4. Results

Figure 2 provides the normalised mean typicality judgements and the proportions of positive responses in the truth value judgement task. The figure suggests that, in general, typicality judgements were less pronounced than the proportions of positive responses in the truth value judgement task. Furthermore, the average typicality values were more evenly distributed across the space of possible answers, whereas the proportions of positive responses in the truth value judgement task clustered around the extremes of 0 and 1. This suggestion is confirmed by a comparison of the variances: the variance in normalised typicality ratings was significantly greater than the variance in the proportions of positive responses in the truth value judgement task ($F(109,109) = 1.9$,

$p < .001$). Both of these observations are captured by the density plot in Figure 3: the modes of the average typicality values are closer to the center than the modes of the proportions of positive responses in the truth value judgement task, and there are more values in the middle region of the space of possible answers in the average typicality values than in the proportions of positive answers in the truth value judgement task.

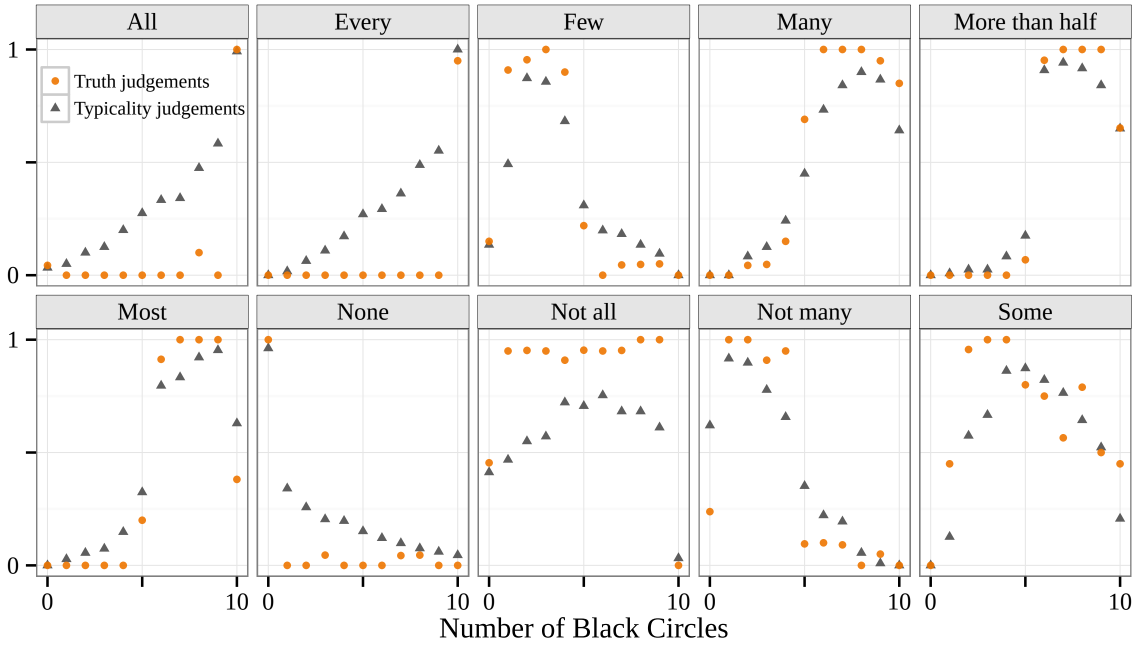


Figure 2: Normalised typicality judgements and proportions of positive responses in the truth value judgement task for ten quantifiers.

One anomalous observation is the proportion of positive responses for ‘some’ in s_8 ($M = .79$). This proportion is unexpectedly higher than in situations with seven ($M = .56$) or nine ($M = .50$) black circles. The difference, however, is not statistically significant in either case. It is presumably caused by the between-participants design of the truth value judgement task: s_8 was judged by different participants than s_7 and s_9. Apparently s_8 was judged by more charitable participants than the other two situations.

Which situations are prototypical of the quantifiers that were investigated? There are at least two ways of answering this question. The first is to take the situations with the highest mean typicality judgements. The second is to take the situations that received the highest typicality judgements from the largest number of participants. For almost all of the quantifiers, these methods lead to the same prototypes. The sole exception is ‘not all’. For this quantifier, the highest mean typicality judgement was for s_6, whereas s_9 was assigned the highest typicality rating by the most participants. This discrepancy reflects a high degree of disagreement between participants in the typicality task for this quantifier. Some participants gave the highest rating to s_0, some to s_9, and

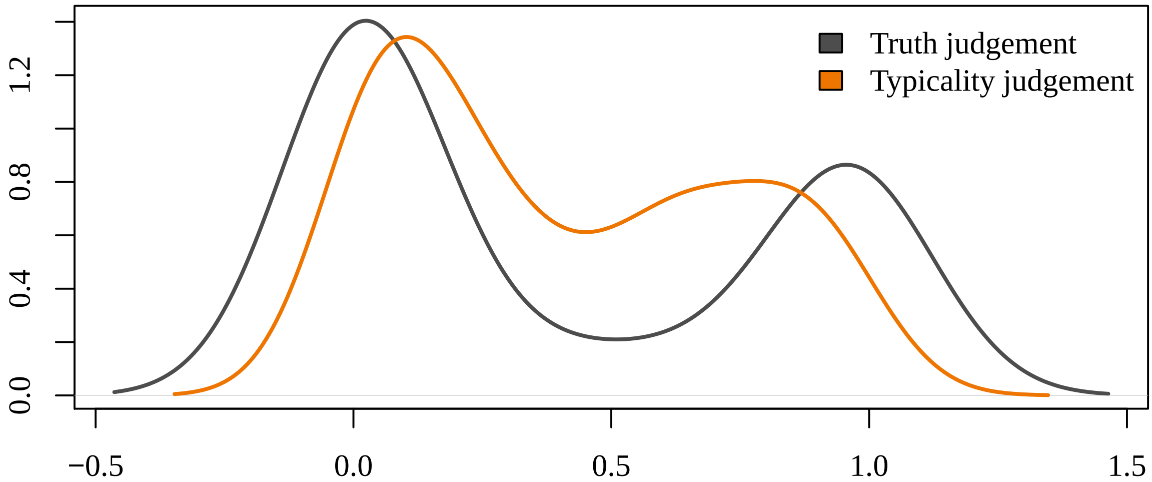


Figure 3: Density plot of the normalised typicality ratings and proportions of positive responses in the truth value judgement task.

some to situations inbetween these extremes. Note that a similar lack of agreement is not visible in the results of the truth value judgement task.

As hypothesised, there is a strong effect of pragmatic inferences on both truth value and typicality judgements. We compared the results for situations where a pragmatic inference was violated with the results for the nearest situation where this was not the case. The mean difference between these situations was higher in the truth value judgement task ($M = .46$) than in the typicality task ($M = .28$). This difference was marginally significant ($t(7) = 1.87$, one-sided $p = .05$). Focusing on the results of the truth value judgement task, the difference was higher for negative quantifiers ($M = .69$) than for positive quantifiers ($M = .32$, $t(6) = 2.46$, one-sided $p = .03$). There was no analogous effect of monotonicity on the effect of pragmatic inferences in the average typicality judgements.

What factors underlie truth value judgements in this experiment? Based on the discussion in Section 2, two possible answers suggest themselves: truth value judgements are determined either by typicality judgements or by set-theoretic definitions. In order to decide between these two possible answers, we constructed three models predicting proportions of positive responses in the truth value judgement task. The first model used the set-theoretic truth definitions given in Table 2 as predictor variable. The second and third model were based on the normalised mean typicality ratings. The second model mapped these typicality ratings straight onto proportions of positive responses in the truth value judgement task. The intuition that underlies this model is that the typicality of an object reflects the likelihood that the sentence is considered true in that situation. The third model dichotomised the typicality judgements on the basis of a cutoff point. We calculated that the optimal cutoff point was 4.4. Values below the cutoff point were mapped to 0 and values above it to 1. According to this model, listeners make truth value judgements by dichotomising their typicality judgements. Unlike the typicality model, it is not obvious what the rationale behind the dichotomised model is. Why do listeners dichotomise typicality judgements on the basis of an apparently arbitrary cut-off point? One possible answer is that this cut-off point reflects whether the sentence is true according to its set-theoretic definition. But in that case the model draws upon set-theoretic definitions just like the set-theoretic model. If this model is to be a competitor to the set-theoretic model, it thus stands in need of a principled motivation.

The mean distances between these three models and the results from the truth value judgement task are given in Table 3. For the set-theoretic model, the mean distances range from .00 for 'every' to .25 for 'some'. For the typicality model, the mean distances range from .04 for 'more than half' to .27 for 'not all'. For the dichotomised typicality model, the mean distances range from .00 for 'every' to .24 for 'not all'.

Quantifier	***Set***	***Typ***	***Dich***	***Quantifier***	***Set***	***Typ***	***Dich***
All	.01	.22	.10	Most	.08	.15	.13
Every	.00	.22	.00	None	.01	.13	.01
Few	.13	.13	.14	Some	.25	.18	.23
Many	.06	.12	.10	Not all	.08	.27	.24
More than half	.04	.04	.04	Not many	.11	.18	.11

Table 3: Mean difference between the proportions of positive responses in the truth value judgement task and (i) the set-theoretic definitions in Table 2 (= ***Set***), (ii) the normalised typicality judgements (= ***Typ***), and (iii) the dichotomised typicality judgements (= ***Dich***)

We compared the distances between these models and the proportions of positive responses in the truth value judgement task. The mean distance was significantly greater for the typicality model (M = .16) than for the set-theoretic model (M = .08, $t(109)$ = 3.77, $p <$ .001). It was also significantly greater for the typicality model than for the dichotomised typicality model (M = .11, $t(109)$ = 2.42, p = .002). The difference in mean distances between the set-theoretic and dichotomised typicality models was marginally significant ($t(109)$ = 1.52, one-sided p = .07).

A substantial part of the lack of fit in all of the models is caused by the confounding effect of pragmatic inferences. Therefore we also used the models to predict a restricted set of the data points that were not affected by pragmatic inferences. This improved all of the models. We once again compared the distances between these limited models and the proportions of positive responses in the truth value judgement task. The mean distance was significantly greater for the typicality model (M = .15) than for the set-theoretic model (M = .03, $t(96)$ = 8.27, $p <$ .001) and for the dichotomised typicality model (M = .06, $t(96)$ = 5.01, $p <$ .001). The mean distance for the dichtomised typicality model was significantly greater than for the set-theoretic model ($t(96)$ = 1.94, one-sided p = .03).

Proportions of positive responses in the truth-value judgement task are thus better approximated by set-theoretic definitions or dichotomised typicality values than by simple typicality values. More-over, there is some evidence that the set-theoretic model is more appropriate than the dichotomised model: it is a marginally better predictor of the proportions of positive responses including data points that are influenced by pragmatic inferences and a significantly better predictor of the proportions of positive responses excluding those data points. In addition, the dichotomised model lacks a principled explanation for the use of a seemingly arbitrary cut-off point in the mean typicality ratings, and the sole plausible motivation seems to invoke set-theoretic truth conditions.

What factors underlie typicality judgements in this experiment? Based on the discussion in Section 2, two possible answers suggest themselves: typicality judgements are determined either by set-theoretic definitions and distance from the prototype or by distance from the prototype alone. In order to decide between these possible answers, we constructed two models predicting mean typicality judgements. The first model included the set-theoretic definitions from Table 2 and distance from the prototype as predictor variables, whereas the second model included distance from the prototype alone. Before these models can be operationalised, however, a number of parameters have to be set. First, what are the prototypes associated with quantified sentences? Second, how to operationalise distance from the prototype? Third, what is the relative importance of set-theoretic definitions and distance from the prototype in the first model? We discuss these issues in turn.

What are the prototypes associated with quantified sentences? As noted before, quantifiers differ in how unambiguous and salient the prototype is: for quantifiers like ‘every’ and ‘none’, all participants converged on the same prototype, for quantifiers like ‘some’ and ‘most’, there was an overall consensus but judgments were not fully unanimous, and for ‘not all’ there was a large amount of disagreement among participants. The choice of prototype seems to be determined by at least two factors: set-theoretic truth conditions and competing quantifiers. To start with the first factor, prototypes are always situations where the sentence is true according to its set-theoretic truth definition. For some quantifiers, however, this still leaves a number of situations to choose from. In that case, the choice might be influenced by competing quantifiers: a prototypical situation is one that is maximally distinct from the prototypical situations of competing quantifiers. This criterion can explain the different choices of prototypes for ‘not all’. Participants who assume that ‘not all’ competes with ‘all’ will consider s_0 as the prototype because that situation is maximally distinct from the prototype for ‘all’. Participants who also took into consideration ‘none’ as a competitor would opt for a situation with around five black circles. Lastly, participants who also considered ‘not many’ might have converged on a prototype that lies somewhere on the upper end of the scale.

This discussion leaves open some further questions. What determines the choice of competing quantifiers? Which quantifiers are in principle available as competitors? How do the alternatives that determine the choice of prototype relate to the alternatives that are involved in the computation of pragmatic inferences? While we believe these questions are interesting and might warrant further analysis, we will simply stipulate that the prototypes are the situations that received the highest mean typicality ratings.

A second question is how to operationalise distance from the prototype. This question has a straightforward answer: the distance between a prototypical situation s_n containing n black circles and another situation s_i equals the absolute difference between n and i.

The final question involves the relative importance of set-theoretic truth conditions and distance from the prototype in the model that included both of these factors. Because we do not have specific expectations about these parameters, we simulated them by means of a Monte Carlo procedure. To this end, we assigned 5,000 random values to both parameters. For each pair of values, we calculated the predicted typicality values and the correlation between these predicted values and the attested typicality values. In the optimal situation, the effect of set-theoretic truth conditions was more than seven times as large as the effect of increasing the distance from the prototype by one. We therefore weighed the two factors accordingly in the first model.

Figure 4 provides a visual overview of the goodness of fit of both models. The correlation between the typicality values predicted by the first model containing both set-theoretic definitions and distance from the prototype, and the attested typicality values was $r = .94$. The correlation between the typicality values predicted by the second model consisting of distance from the prototype alone and the attested typicality values was $r = .83$. We compared the two models on the basis of the absolute differences between predicted and attested typicality values. The mean difference was significantly higher for the second model ($M = .21$) than for the first model ($M = .12$, $t(109) = 6.53$, $p < .001$). As before, the fit of both models is relatively poor for situations that are excluded for pragmatic reasons. Excluding those data points from consideration leads to correlations of $r = .96$ for the first model and $r = .87$ for the second one.

What do these results tell us about the questions we posed at the end of the Introduction? What is the relationship between set-theoretic truth conditions and typicality judgements? Do listeners have access to set-theoretic definitions of quantified sentences? In the following section, we discuss these questions on the basis of the foregoing results.

4. General discussion

In the Introduction, we observed that the interpretation of quantified statements is often finer-grained than what is encoded in their set-theoretic definitions. For example, according to its set-theoretic truth conditions, the statement ‘Some A are not B’ is true in all situations where not all of the A are B. But when participants are asked how many A are B given that this statement is true, Newstead et al. (1987) found that their estimates are much more precise and range between 55% and 89% of the total number of A.

These findings can be modelled in different ways. Some authors have proposed that quantifiers denote probability distributions over situations. This proposal, however, presupposes knowledge about the total set size, whereas it seems possible to interpret quantifiers even in the absence of this knowledge. The current findings provide a further argument against a probabilistic view on

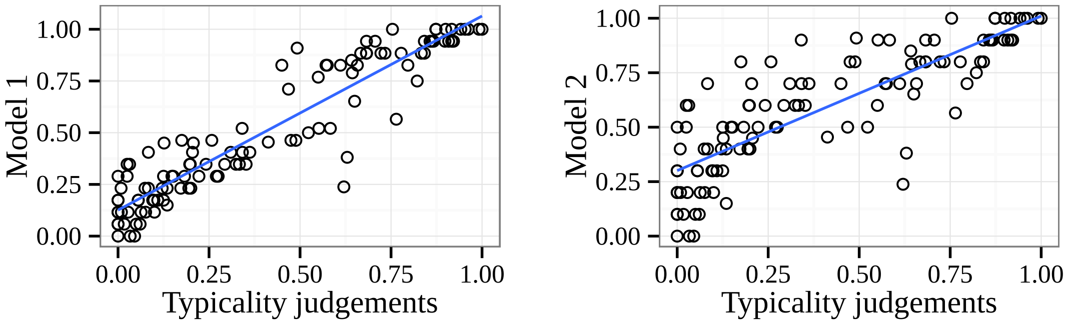


Figure 4: Scatterplot with the typicality values predicted by the two models and the typicality judgements found in the experiment.

quantification. Consider the typicality judgements for the quantifiers 'all', 'every', and 'none'. These steadily decrease with the distance from the prototypical situation, which means that all situations except for the most distant one received a positive rating. It seems implausible, however, to conclude from these findings that, for example, that given that 'All the circles are black', the probability of a situation with three black circles is anything other than zero.

In order to avoid these issues, we have modelled the finer-grained interpretation of quantifiers by assuming that they denote typicality functions. In that case, is it still plausible to suppose that listeners have access to set-theoretic definitions of quantifiers? The results of our experiments provide a number of arguments to substantiate the role of set-theoretic definitions in the interpretation of quantifiers. First, we constructed three models to predict proportions of positive responses in the truth value judgement task. One model consisted of set-theoretic definitions; the second model consisted of typicality judgements; and the third model dichotomised these typicality judgements based on a cutoff point. We found that the absolute deviations from the first model were significantly smaller than those from the second model, and marginally smaller than those from the third model.

These findings demonstrate that the typicality of a situation should not be equated to the likelihood that the corresponding quantified statement is judged true. The evidence against the view that truth value judgements are dichotomised typicality judgements, however, was less convincing. Still, even if we assumed that truth value judgements are dichotomised typicality judgements, it wouldn't be clear *why* this should be so, and set-theoretic definitions would still have to feature in the explanation of the typicality judgements themselves.

We constructed two models to predict typicality judgements. The first model consisted of set-theoretic definitions and distance from the prototype, where the former factor was given seven times as much weight as the latter. The second model consisted of distance from the prototype alone. We found that absolute deviations from the first model were significantly smaller than those from the second model. This finding again indicates that listeners have access to set-theoretic definitions of quantifiers.

A related argument in favour of the view that set-theoretic definitions feature in the interpretation of quantified statements is the rarity of disagreement between participants about whether a quantified statement was true or false. More precisely, aside from situations where pragmatic inferences play a role, there were four proportions of positive responses in the range between 0.1 and 0.9. Three of these involved the s_5 situation for the proportional quantifiers 'most', 'many', and 'few', and one involved the s_4 situation for 'many'. These three quantifiers all involve some kind of threshold value. So a possible explanation for these anomalous observations is that participants might have disagreed about the exact value of this threshold. By contrast, there was a substantial amount of disagreement in the typicality judgements for all quantifiers. If truth value judgements were calculated on the basis of typicality judgements, we would have expected a similar amount of disagreement in both of these measures.

Furthermore, it might be argued that set-theoretic definitions are necessary to account for the interpretation of embedded quantifiers. For example, it seems inevitable to suppose that the interpretation of 'not all' is a function of the interpretation of 'all'. This relationship is apparent in the truth value judgements but not in the typicality judgements: the correlation between the proportions of positive responses in the truth value judgements task for 'all' and 'not all' is -.88 and, if the effect of the pragmatic inference for 'not all' is taken into account, -.98. This strength of association is absent in the typicality judgements: the correlation between typicality judgements for 'all' and 'not all' is -.47. (Note, however, that the typicality structure of 'many' does provide a reliable fit to the typicality structure of 'not many'.)

We have thus provided a number of arguments in favour of the view that listeners have access to set-theoretic interpretations of quantifiers. One apparent exception to this rule is 'some'. The set-theoretic definition for this quantifier was a poor fit to the actual truth value judgements. Truth value judgements were equivocal for all situations other than those with zero black circles, where it was judged false by all participants, and those with two, three, or four black circles, where it was judged true by all or almost all participants. As noted, however, 'some' is exceptional in that it licenses three pragmatic inferences. Surprisingly, participants also make distinctions between situations that do not correspond to a pragmatic inference: for example, the statements was judged true by more participants in s_6 than in s_9, even though both situations violate the pragmatic inference that not most of the circles are black. Apparently, then, distance from the prototype also influences judgements about what to do with statements that violate a pragmatic inference: the likelihood that a sentence is judged false in such a situation depends on how salient its violation is.

Aside from demonstrating the role of set-theoretic definitions in the interpretation of quantifiers, we have provided a model for truth value and typicality judgements for quantified statements based on set-theoretic definitions, distance from the prototype, and pragmatic inferences. The interpretation of quantifiers is thus a multidimensional phenomenon that warrants further investigation.

References

Barwise, J. and R. Cooper (1981). Generalized quantifiers and natural language. *Linguistics & Philosophy 4*(2), 159–219.

Bott, L. and I. A. Noveck (2004). Some utterances are underinformative: the onset and time course of scalar inferences. *Journal of Memory and Language 51*(3), 437–457.

Chase, C. I. (1969). Often is where you find it. *American Psychologist 24*(11), 1043.

Kamp, H. and B. Partee (1995). Prototype theory and compositionality. *Cognition 57*(2), 129–191.

Katsos, N. and D. V. M. Bishop (2010). Pragmatic tolerance: implications for the acquisition of informativeness and implicature. *Cognition 20*(1), 67–81.

Keenan, E. L. and J. Stavi (1986). A semantic characterization of natural language determiners. *Linguistics and Philosophy 9*(3), 253–326.

Lakoff, G. (1973). Hedges: a study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic 2*(4), 458–508.

McCloskey, M. E. and S. Glucksberg (1978). Natural categories: well defined or fuzzy sets? *Memory & Cognition 6*(4), 462–472.

Montague, R. (1973). The proper treatment of quantification in ordinary English. In K. J. J. Hintikka, J. M. E. Moravcsik, and P. Suppes (Eds.), *Approaches to natural language*, pp. 221–242. Dordrecht: Reidel.

Moxey, L. M. and A. J. Sanford (1993). Prior expectations and the interpretation of natural language quantifiers. *European Journal of Cognitive Psychology 5*(1), 73–91.

Newstead, S. E., P. Pollard, and D. Riezebos (1987). The effect of set size on the interpretation of quantifiers used in rating scales. *Applied Ergonomics 18*(3), 178–182.

Pepper, S. and L. S. Prytulak (1974). Sometimes frequently means seldom: context effects in the interpretation of quantitative expressions. *Journal of Research in Personality 8*(1), 95–101.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology 104*(3), 192–233.

Rosch, E. and C. B. Mervis (1975). Family resemblances: studies in the internal structure of categories. *Cognitive Psychology 7*(4), 573–605.

Spector, B. (2006). *Aspects de la pragmatique des opérateurs logiques*. Ph. D. thesis, Université de Paris VII.

Yildirim, I., J. Degen, M. K. Tanenhaus, and T. F. Jaeger (2013). Linguistic variability and adaptation in quantifier meanings. In M. Knauff, M. Pauen, N. Sebanz, and I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, pp. 3835–3840. Austin, TX: Cognitive Science Society.

Zevakhina, N. (2012). Strength and similarity of scalar alternatives. In A. Aguilar Guevara, A. Chernilovskaya, and R. Nouwen (Eds.), *Proceedings of Sinn und Bedeutung 16*, pp. 647–658. MIT Working Papers in Linguistics.