How many *many*s? Exploring semantic theories with data-driven computational models[1]

Anthea Schöller — University of Tübingen
Michael Franke — University of Tübingen

Abstract. We use a data-driven computational inference approach to address the question whether it is plausible to maintain that there is a stable core semantics that governs the interpretation of cardinal and proportional *many* across different contexts. Adopting the idea that the denotation of *many* is a function of a stable threshold parameter that applies to a contextually-variable probability distribution that captures prior expectations, we demonstrate that it is possible to maintain that there is a single fixed threshold for *many*'s cardinal and proportional use, although models that allow for non-uniform thresholds or lexical ambiguity may have a slightly better empirical fit to our data.

Keywords: *many*, quantifiers, ambiguity, computational modeling, experimental data, context

1. Introduction

How do speakers use vague expressions like *many, few, tall* or *good*? What is their relation to the context and how does a learner acquire this knowledge? Assuming that language learning is economical and efficient, it is plausible that vague expressions have a stable core meaning which determines their use in any context. The opposing view would be that these words' meanings differ in each context. The second assumption makes very implausible predictions, however, namely that learners need to master the use of vague expressions anew for each context. Here, we argue for the first hypothesis and explore the relationship between vague expressions and the context.

How to capture the context-dependence of vague expressions, is a challenge to linguistic theory. In this paper, we will focus on *few* and *many*, which, similar to gradable adjectives like *tall* or *expensive*, express a number, or, in more abstract terms, a degree in a vague manner. It is hard to pin down a precise denotation in each context and there is ample variance between contexts, as exemplified in (1).

(1) a. Few of Martha's grandchildren could afford to buy a car when turning 18.
b. Few US citizens went to the polls in the last elections.

For a long time, there has been a debate in the literature about how these expression's vagueness and their interaction with the context can be captured in their semantics (Hörmann 1983, Partee

[1]We thank Fabian Dablander for practical assistance and audiences in Tübingen, Stanford and Göttingen for their feedback on this work. MF is supported by the Institutional Strategy of the University of Tübingen (Deutsche Forschungsgemeinschaft, ZUK 63). Both authors gratefully acknowledge support by Priority Program XPrag.de (DFG Schwerpunktprogramm 1727).

1989, Clark 1991, Fernando and Kamp 1996, Solt 2009). There is even dispute about how to classify them. They are labeled "quantifiers" (Barwise and Cooper 1981), "scalar quantifiers" (Hackl 2000) or "adjectives of quantity" (Solt 2009). Furthermore, some authors have questioned whether one lexical entry is sufficient: Partee (1989) proposed that *few* and *many* are ambiguous between a *cardinal* and a *proportional* reading (more below). But hardly any of these theories makes concrete predictions about which and how particular contextual parameters fix or constrain interpretation.

Previous related work (Schoeller and Franke 2015) looked at experimental data on production and interpretation of *few* and *many* and applied a data-driven computational model to investigate whether it is possible to maintain that there is a common fixed semantic core meaning that plausibly explains proportional usages across a number of diverse contexts. The idea underlying the computational model was that of Clark (1991) and Fernando and Kamp (1996): truth conditions of *few* and *many* are a function of a fixed threshold, a context-independent meaning component, on the probability density function of a distribution that captures prior expectations (more on this below). The central question of this paper is whether the findings from Schoeller and Franke (2015) extend to proportional readings of *many*.[2] Does it have a stable core meaning and is even a unified account of both readings possible? Or are we dealing with a genuine lexical ambiguity?

Section 2 introduces relevant background. Section 3 describes how we experimentally gathered data and used statistical analyses to learn about the context-sensitivity of proportional *many.* Section 4 explains how we can turn a semantic theory into a computational model of language use. Sections 5 and 6 describe more experiments to gather data which was used in computational modeling (Section 7) to see whether a unified treatment of cardinal and proportional readings is possible. By doing so, we want to contribute to the discussion about the ambiguity of *many* and the interaction between semantics and the context and follow the recent trend of combining theoretical linguistics, experimental data and computational modeling. Section 8 concludes with a methodological reflection.

2. Semantic Background

Partee (1989) argued that *few* and *many* can be read in two ways:[3]

(2) Cardinal reading of "Few/Many As are B"
   a. Few: $|A \cap B| \leq x_{max}$
   b. Many: $|A \cap B| \geq x_{min}$

(3) Proportional reading of "Few/Many As are B"
   a. *Few*: $\frac{|A \cap B|}{|A|} \leq k_{max}$
   b. *Many*: $\frac{|A \cap B|}{|A|} \geq k_{min}$

Partee (1989) suggests that the quantifiers' cardinal reading has a meaning "like that of the cardinal

[2] We focus on *many* because we want to sidestep additional complications that seem involved in the use of *few*.
[3] Italicized A/B is the extension of predicate A/B.

numbers, *at least* x_{min}, with the vagueness located in the unspecified choice of x_{min} ... The cardinal reading of *few* is similar except that it means *at most* x_{max}, and x_{max} is generally understood to be small" (Partee 1989: p.1)[4] This theory is intuitively appealing, but the threshold parameters x_{min}, x_{max}, k_{min} and k_{max} are not specified and it is not clear how their value changes across contexts. In Schoeller and Franke (2015), we focused on the *cardinal surprise reading* as in (4) and described by Clark (1991) and Fernando and Kamp (1996).

(4) a. Joe eats many burgers. ($\rightsquigarrow$ Joe eats more burgers than expected of him.)
b. Melanie owns many pairs of shoes. ($\rightsquigarrow$ Melanie owns more than expected of her.)

We experimentally investigated the production and interpretation of cardinal *few* and *many* and found that it is at least plausible that the relationship between the numerical denotation and the context can be captured by a fixed semantic parameter that interacts with contextually variable prior expectations (e.g., about the number of burgers a guy like Joe can be expected to eat). The semantic predictions and the model we applied are laid out in Section 4.

In this paper, we want to investigate the interaction between the context and proportional *many.* Partee (1989) discusses the proportional reading of *few* and *many* in sentences like (5) with the semantics in (3).

(5) a. Many of the US citizens live in big cities.
b. Few of the US citizens speak German.

Sentence (5a) is true if a large proportion of US citizens live in big cities; at least *k*. "We may think of *k* either as a fraction between 0 and 1 or as a percentage" (Partee 1989: p. 2). For *few*, sentence (5d) is true if a small proportion of US citizens speaks German, at least *k*. How to define the size of the fraction *k* which determines of usage of *few* and *many* is left unspecified, however. Furthermore, (3) does not tell us what the influence of the context on threshold *k* is or whether it is assumed to be a fixed proportion. In an experiment on the interpretation of proportional *many*, we want to find out whether it is possible to define *k* independently of the context.

3. Experiment: Influence of the context on the interpretation of proportional *many*

In this experiment on the interpretation of sentences with proportional *many*, we want to investigate the influence of the contextual expectations. Furthermore, we want to find out whether it makes a difference to use *many* in the plain form "many" or in the partitive construction "many of the" and whether the number of objects in the context influences the interpretation.

[4]Partee (1989) labels both variables with n. For consistency with the theory proposed in Section 2 we use x_{max} and x_{min} instead.

3.1. Methods and material

We ran an experiment on Amazon's Mechanical Turk and elicited data from 160 participants for a reimbursement of 0.50$. Participants who are not self-reported native speakers of English or showed clearly uncooperative behaviour were excluded. At the beginning of the experiment, each participant was randomly assigned to condition [-/+ partitive]. [-partitive] means, that every sentence was presented with plain "many", whereas in the [+partitive] condition "many of the" was used. Every participant saw 16 items. A sentence introduced the context and the amount of the objects under discussion. Each item was paired with two numbers of the form {N, 3/4N} and one of these numbers was randomly chosen in each trial as the total amount. A sentence containing the quantifier was randomly chosen from two conditions [HP/LP], high probability or low probability. The two conditions differed in the comparison class set in the relative clause. We set the comparison classes in a way that we expect higher answers in high probability contexts. We made sure that the two relative clauses per item are a minimal pair. Most of them differed only in contrasting adjectives. A sample item is given in (6) and (7). In a free production task, participants were asked to guess the number that they think "many" or "many of the" refers to.

(6) There were **9/12** muffins on the kitchen table in Eds flat.

HP: Ed, who arrived feeling hungry, ate **many/many of the** muffins.

LP: Ed, who arrived feeling full, ate **many/many of the** muffins.

How **many/many of the** muffins do you think Ed ate?

(7) When moving flat, Martha packed **15/20** big boxes.

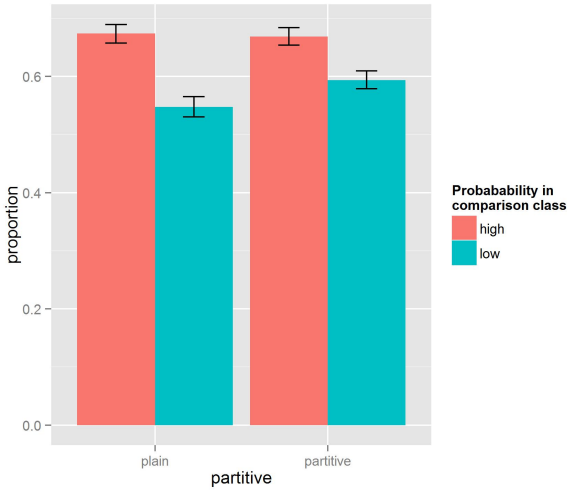
HP: Martha, who is a strong woman, carried **many/many of the** boxes herself.

LP: Martha, who is a weak woman, carried **many/many of the** boxes herself.

How **many/many of the** boxes do you think Martha carried?

3.2. Hypotheses

We expect that the comparison class has an effect on the interpretation of "many". We expect that people interpret *many* as higher numbers / proportions in the high probability condition than in the low probability condition. The partitive construction should facilitate a proportional reading. This is why we expect less of an influence of the comparison class in sentences with "many of the". The difference between low and high probability should not be as big as for the sentences with plain "many". Furthermore, a pre-study suggests that the number of objects in the context influences the interpretation of "many". We expect that if the number is high, it is more likely that the proportional interpretation is lower than if participants are presented with a low number of objects. However, as the range of amount is not very big in this experiment, it would not be surprising to find no effect of amount.



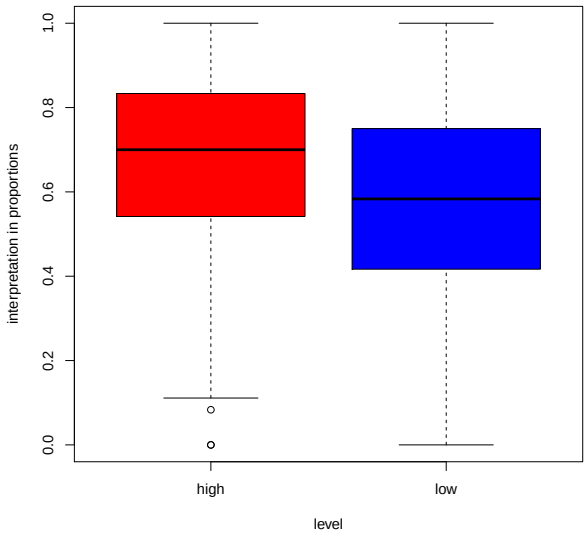
Condition	Mean proportion
plain & low	0.58
plain & high	0.73
partitive & low	0.58
partitive & high	0.67

Figure 1: Mean ratings for the interpretation of *many* in proportions of *N* for both high and low probability contexts and without or with the partitive construction

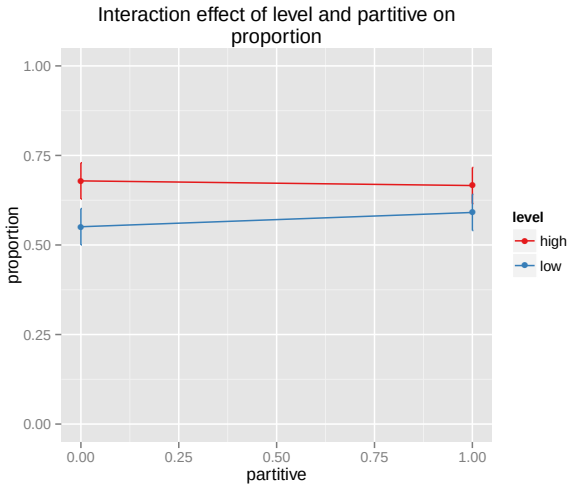
3.3. Results

Figure 1 gives a first impression of the outcome of the experiment. When looking at the mean proportions of *N* that were given as the interpretation of *many*, we see a clear difference between LP and HP condition. This is a first piece of evidence which supports the hypothesis that prior expectations influence interpretation. Furthermore, the difference between low and high probability is greater in the plain condition than when the partitive is used. Whether these differences are statistically significant will be analyzed in the following.

At first we specified a mixed linear effects regression model predicting proportional interpretations for "many" which included the main effects "level" (high or low probability sentence), "amount" (number in context), "partitive" (plain or partitive "many") and an interaction of these three predictors. In terms of random effects, the initial model had the maximal random effects structure as justified by the design (Bates et al. 2013). We removed redundant random effects by running a principle component analysis and arrived at a parsimonious model (Bates et al. 2015). The final model included both varying intercepts for "participant" and "item", as well as a random "participant" slope for "level". In terms of the fixed effects, only "level" was included as a main effect. We found that participants gave significantly lower ratings in the low-level condition ($\beta = -0.128, SE = 0.013, p < 0.001$). Figure 2a shows the predicted interpretation of *many* in proportions of *N* in both HP and LP condition of the factor "level". The data suggests that participants interpret *many* as a lower proportion of *N* when it is presented in a low probability context than when *many* occurs in a high probability context. We can interpret the fact that the factor "level" was identified as a main effect as evidence that the context influences the interpretation of proportional *many*. This effect was modulated by an interaction of "level" with "partitive"



(a) Main effect of level



(b) Interaction between level and partitive

Figure 2: Differences in proportional ratings between high and low probability contexts and with (partitive = 1) or without (partitive = 0) partitive construction.

($\beta = -0.052, SE = 0.018, p < 0.005$). Figure 2b shows again that ratings are lower in the low-level condition, for both forms of the quantifier. However, in the low-level condition the partitive construction (x-axis: partitive = 1) triggers higher ratings than plain "many" (x-axis: partitive = 0). This slope is reversed in the high-level condition. The plot shows that the factor "partitive" has an effect on the interpretation of *many* in that it interacts with "level". However, the plot also shows that "partitive" is not a strong enough fixed factor to qualify as a main effect. In Subsection 3.2 we hypothesized that the partitive construction should facilitate a proportional reading and allow less of a difference between the two context conditions. Figure 2b displays that the difference between low and high probability contexts is slightly bigger in sentences without a partitive but that the difference to sentences with a partitive construction is not significant.

3.4. Discussion

The linear mixed effects regression suggests that the comparison class has a significant effect on the interpretation of *many*. This contradicts a theory which assumes one fixed value for the proportion *k*. Rather, the semantics should comprise *many*'s interaction with the context. Interestingly, neither the factor "amount" nor the factor "partitive" were significant. That the use of these two factors does not make a difference leaves open the possibility of a unified semantics because cardinal *many* cannot be combined with the partitive nor is its range restricted by an upper bound.

As a next step, we want to examine more closely how the interpretation of proportional *many* is affected by the context. To do this, we want to measure people's prior expectations of typical amounts in the contexts we used and apply a computational model to our data which formalizes a particular way of mapping these contextual expectations onto predictions about language use. The next section introduces this model.

4. CFK semantics and computational modeling

A concrete proposal of how contextual expectations map onto truth conditions of *many* and *few*, was first suggested tentatively by Clark (1991) and formally spelled out by Fernando and Kamp (1996). We will call it the Clark-Fernando-Kamp (CFK) semantics. The CFK semantics describes the reading of *few* and *many* in sentences like (8) as the cardinal surprise reading because it treats it as intensional and expresses that the number in question is lower or higher than expected.

(8) Melanie owns few/many pairs of shoes.

(9) **CFK Semantics**

a. [[Few A are B]] = 1 iff $|A \cap B| \leq x_{max}$
where $x_{max} = \max\{n \in \mathbb{N} \mid P(|A \cap B| \leq n) < \theta_{few}\}$

b. [[Many A are B]] iff $|A \cap B| \geq x_{min}$
where $x_{min} = \min\{n \in \mathbb{N} \mid P(|A \cap B| \leq n) > \theta_{many}\}$

The CFK semantics in (9) aims to explain the contextually variable thresholds x_{max} and x_{min} from the truth-conditions in (2) as a function of prior expectations *P* and a pair of fixed thresholds θ_{few} and θ_{many} on the cumulative distribution derived from *P*. Thresholds θ_{few} and θ_{many} can then be conceived of as the contextually-stable semantic core meaning of *many* and *few* that would help explain how vague quantifiers can be meaningfully used and faithfully acquired. Applied to example (8), given (9b) the sentence is true if the number of shoes owned by Melanie is greater than x_{min}. In turn, x_{min} is specified as the lowest number for which the cumulative density mass of the prior expectation *P* over numbers of shoes that Melanie owns is higher than the semantically fixed threshold θ_{many}.

The CFK semantics looks intuitively appealing, but how can such a proposal even be tested? Toward this end, we look at a computational model. The idea is that we use empirical measures of expectations P (for each relevant context) and feed it into the model. The model then predicts threshold values via (9) and maps these onto a likelihood of judging a statement with *few* or *many* as true in a particular context and a likelihood of interpreting it in a particular way. (see Schoeller and Franke 2015: for details). Here, we focus on the interpretation of *many*.

Data from experiments on the interpretation of *many* is used to "reverse infer" credible values for the threshold θ_{many} by Bayesian inference. Concretely, we compare models in which we infer just

one treshold θ_{many} that applies to all contexts with an alternative model that uses an independent θ_{many} for every context. The question which model better fits the data can then be addressed by statistical model comparison. This fuels the discussion of the theoretical question whether the CFK semantics as a whole and belief in a uniform θ_{many} in particular are plausible assumptions.

5. Experiment: Prior elicitation for proportional *many*

This experiment was designed to gather data about people's prior expectations concerning the contexts used in the interpretation task from Section 3.

5.1. Methods and material

We ran the experiment on Amazon's Mechanical Turk and elicited data from 160 participants for a reimbursement of 0.35$. Only data by native speakers of English was considered. We designed the material in a way that ensured compatibility with the interpretation task. Because the analysis of the interpretation data did not support a significant main effect of the partitive construction (*many* vs. *many of the*) or of number (the number of objects or activities presented), we decided to not further investigate these two factors. However, we included the factor level (low or high probability of the event) because we found that this was a main effect. We only elicited prior expectations of 10 of the previous 16 items. Each item contained a fixed number of objects or activities. Depending on the number introduced in the item, we presented participants with 10, 13 or 16 intervals and asked to rate the probability of this number by adjusting a slider on a scale ranging from "very unlikely" to "extremely likely". Two sample items are given below, the remainder in Appendix A:

(10) There were 12 muffins on the kitchen table in Eds flat.
HP: Ed arrived feeling hungry.
LP: Ed arrived feeling full.
How many of the muffins do you think Ed ate?
{0,1,2,3,4,5,6,7,8,9,10,11,12}

(11) When moving flat, Martha packed 15 big boxes.
HP: Martha is a strong woman.
LP: Martha is a weak woman.
How many of the boxes do you think Martha carried?
{0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15}

5.2. Hypotheses

In the interpretation task we found that the comparison class (e.g., a hungry vs. full person eating muffins) had an effect on the interpretation of *many*. Furthermore, our previous research suggests that for cardinal *few* and *many* production and interpretation can be predicted by a context-independent threshold, which can be formalized as a percentage on the cumulative density mass of the prior expectation (Schoeller and Franke 2015). The findings in Section 3 suggest that it might be plausible to find such a context-independent threshold for the interpretation of proportional *many*, too. Whether the actual percentage of the cumulative density mass of the prior is the same as for cardinal *many* remains to be seen.

5.3. Results

Figure 3 displays the probability distributions we measured and which we take to represent the prior expectations. We first normalized the ratings of each item within participants. Second, we calculated the mean rating of each interval for all participants. These probability distributions are input to the computational model which estimates context-independent threshold values if the data suggests that they exist.

6. Experiment: Cardinal *many*

As a sanity check of the findings in Schoeller and Franke (2015), we reran the experiments on production and interpretation of cardinal *few* and *many* and also elicited prior expectation of the presented contexts. The design remained unchanged (see Schoeller and Franke (2015) for a detailed description), but we replaced some items for which the context or the phrasing was unclear. This data will be used in Section 7 when a computational model is applied to data from both cardinal and proportional uses of *many*. The items can be found in Appendix B.

7. Computational models

The CFK semantics' predictions were transformed into a probabilistic computational model of interpretation behavior. Latent semantic parameters, in this case the threshold values θ_{few} and θ_{many}, are estimated on the basis of the experimental data (see Section 4). To ensure comparability, we estimated the parameters based on interpretation data only.
We will run two or three versions of the model for each data set. The first version follows the predictions of the CFK semantics and estimates **one** threshold value θ_{many} which explains the interpretation of *many* for each item. We call it the general threshold model (**GTM**). The second version captures the alternative hypothesis that there need not be a universally shared threshold. Consequently, we allow for an **individual** threshold parameter for each context. This model is called individual threshold model (**ITM**). It is likely that the ITM yields a good fit to the data;

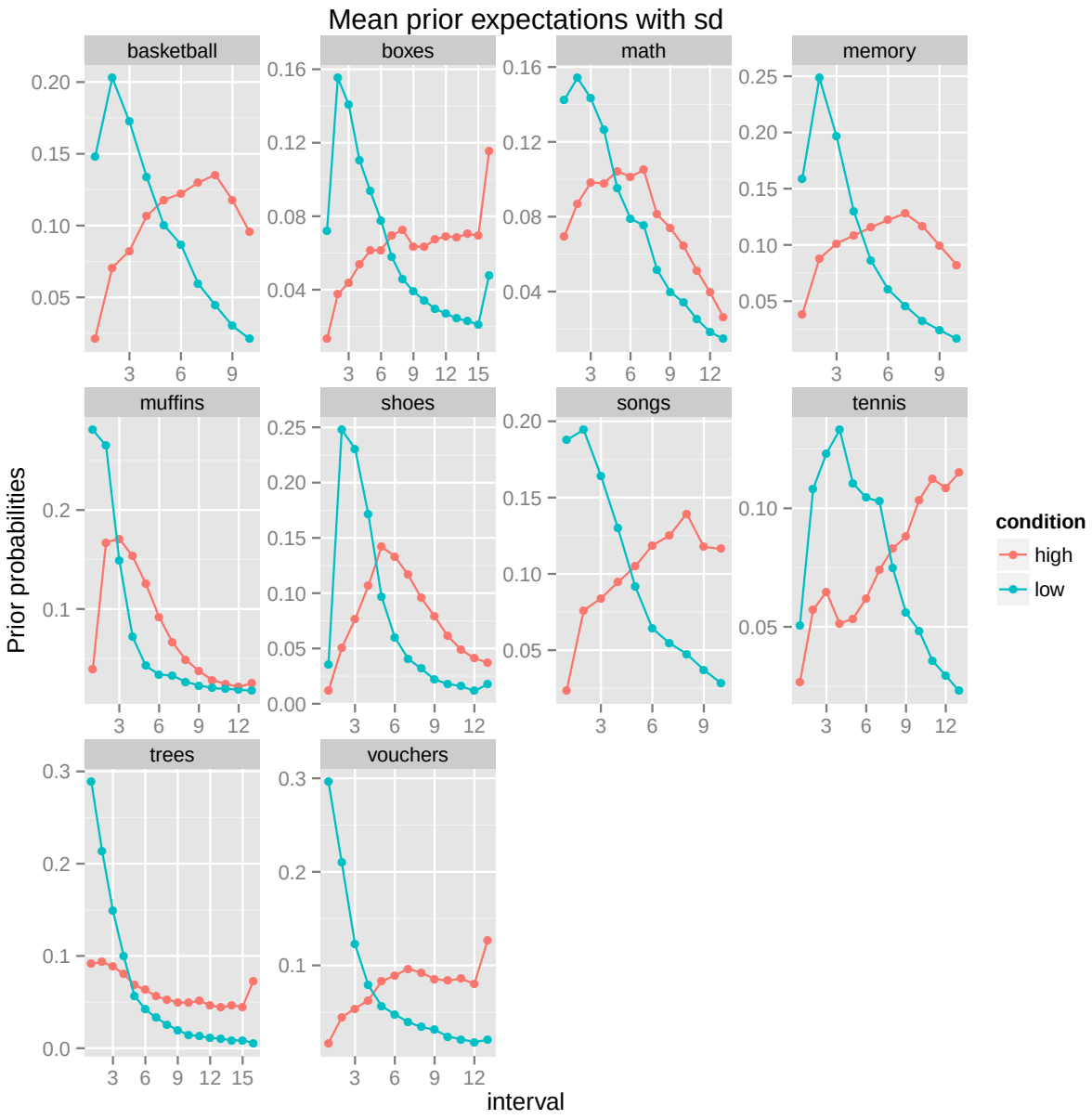


Figure 3: Proportional *many*, prior expectations for both context conditions

however, this flexibility comes at a price: it is much more complex because it is forced to estimate one parameter per context and not one for each of them as the GTM does. The third version of the model, the **threshold per reading** model (**TRM**) tests the hypothesis that *few* and *many* are lexically ambiguous. It is based on the assumption that both readings are captured by the CFK semantics but that their threshold values are different. We will compare the three versions' fit to the data in a statistical model comparison using each model's DIC value. This concept combines a measure of model fit with a measure of model complexity and will be introduced and applied in

Section 7.5. For each version of the model and each data set, we collected 10,000 samples from 2 MCMC chains after a burn-in of 10,000. This ensured convergence for every model, as measured by $\hat{R}$.

For each context, we are not only interested in its DIC value, but also in the variance of the individual thresholds. To check whether these individual thresholds are similar, we will estimate 95% credible intervals for the marginalized posteriors over each threshold in the ITM. A 95% credible interval is, intuitively put, an interval of values that are sufficiently plausible to warrant belief in (see Kruschke 2014). For example, if [0.65;0.75] were the 95% credible interval for θ^i_{many} for some item i, we should be reasonably certain that the true value of θ^i_{many} is in this interval. If the contexts credible parameter values for θ^i_{many} overlap on some interval, this interval is where a uniform semantic threshold might reside.

7.1. Hypothesis:

We expect that the model comparison favors the GTM because it predicts that the denotation of *many* is calculated in the same way for each context. This way, the vagueness of the expression is preserved because its denotation is still dependent on the contextual input, but the procedure is fixed. Most importantly, this is what we expect from the perspective of processing and language learning. We can interpret *many* in infinitely many contexts justs as we are able to understand an infinite number of sentences by compositionality.

7.2. Proportional *many* (20 contexts)

General θ model: The mean of the posterior of θ_{many} was estimated via sampling as 0.83. This model has a DIC = 977.9 and pD = 24.7.

Individual θs model: In total, 12 of 20 thresholds' HDIs overlapped in 0.83, among them 8 contexts in the low probability conditions (see Figure 4). DIC = 889.2 and pD = 55.9.

7.3. Cardinal *many* (14 contexts)

General θ model: The mean of the posterior of θ_{many} was estimated via sampling as 0.69. DIC = 1076.3 and pD = 19.2.

Individual θs model: For cardinal *many*, 7 of 14 items' HDIs overlap in [0.65, 0.70] which includes θ_{many} = 0.69 estimated by the GTM (see Figure 5). DIC = 995.0 and pD = 42.7.

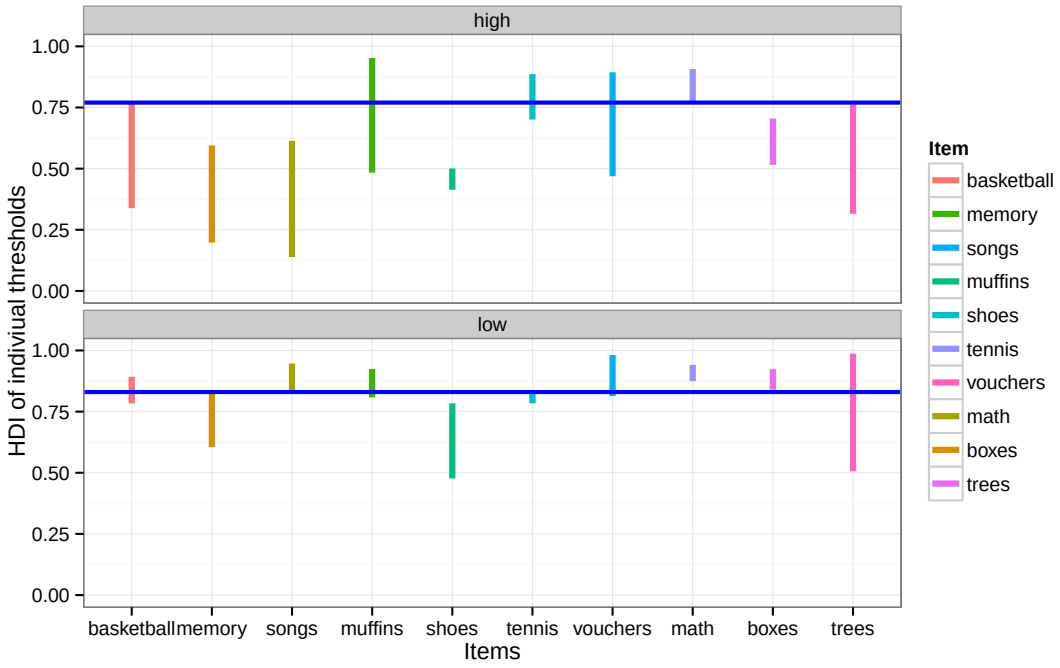


Figure 4: Proportional *many*: posteriors of individual threshold values in both conditions

7.4. Fusing cardinal and proportional readings (34 contexts)

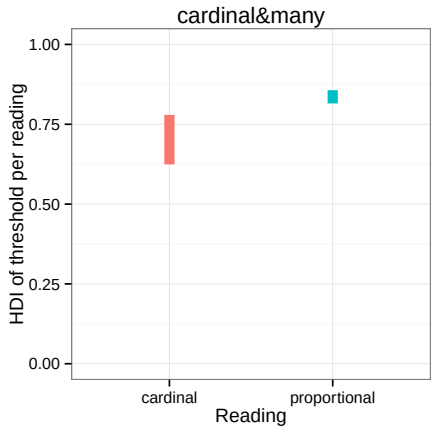


Figure 6: threshold values for cardinal and proportional reading in TRM

General θ model: This model tests whether it is possible to find one threshold which explains both readings of *many* at once. The mean of the posterior of θ_{many} was estimated via sampling as $\theta_{many} = 0.83$. (Note: that the value for a general θ_{many} is this high is very likely based in the fact that 20 of the 34 contexts are proportional, compare Subsections 7.2 and 7.3). DIC = 2061.0 and pD = 39.7.

θ per reading model: This model estimates one threshold per readings. $\theta_{many:prop}$ captures all interpretations of proportional *many* and $\theta_{many:card}$ all interpretations of cardinal *many*. The HDI of the threshold for the proportional uses of *many* is $\theta_{many:prop}$ = [0.82, 0.86] and for the cardinal uses $\theta_{many:card}$ = [0.62, 0.78] (see Figure 6). DIC = 2056.4 and pD = 45.7.

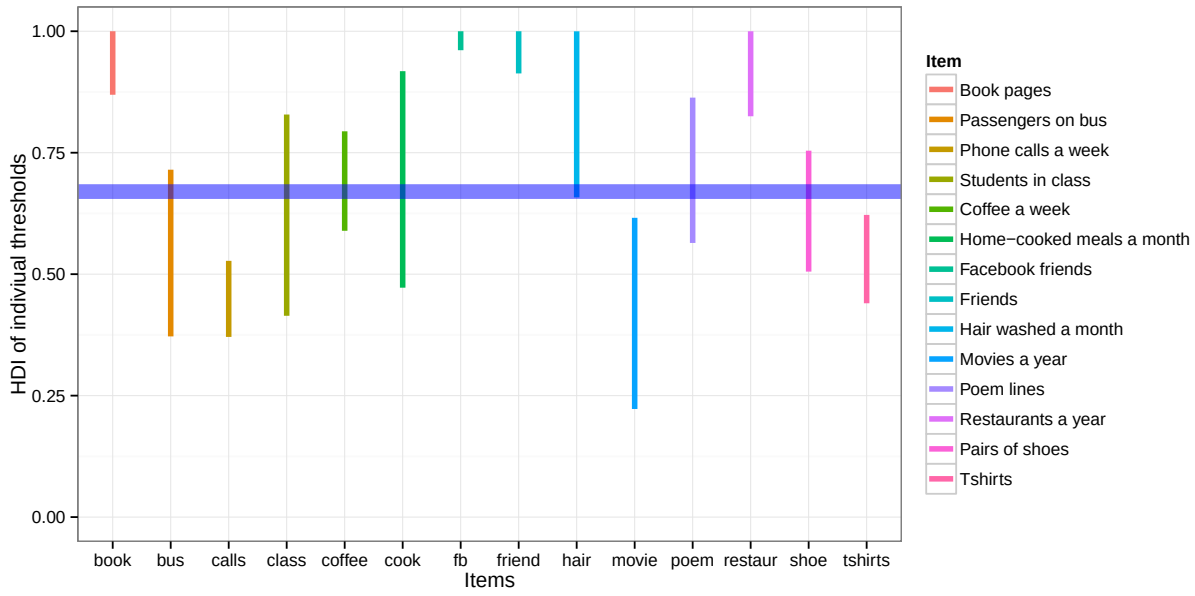
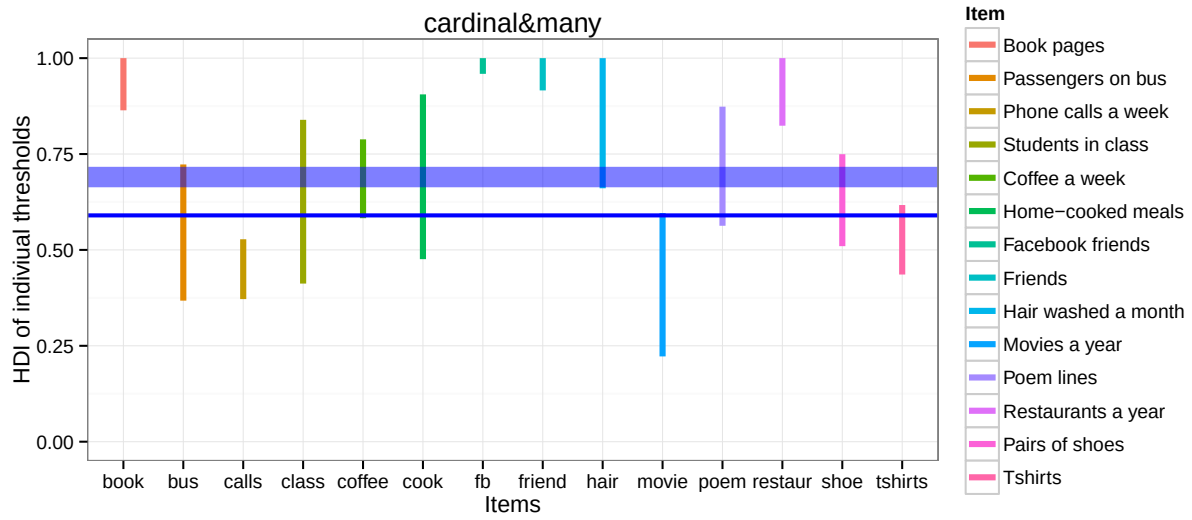


Figure 5: Cardinal *many*: predictions for individual thresholds

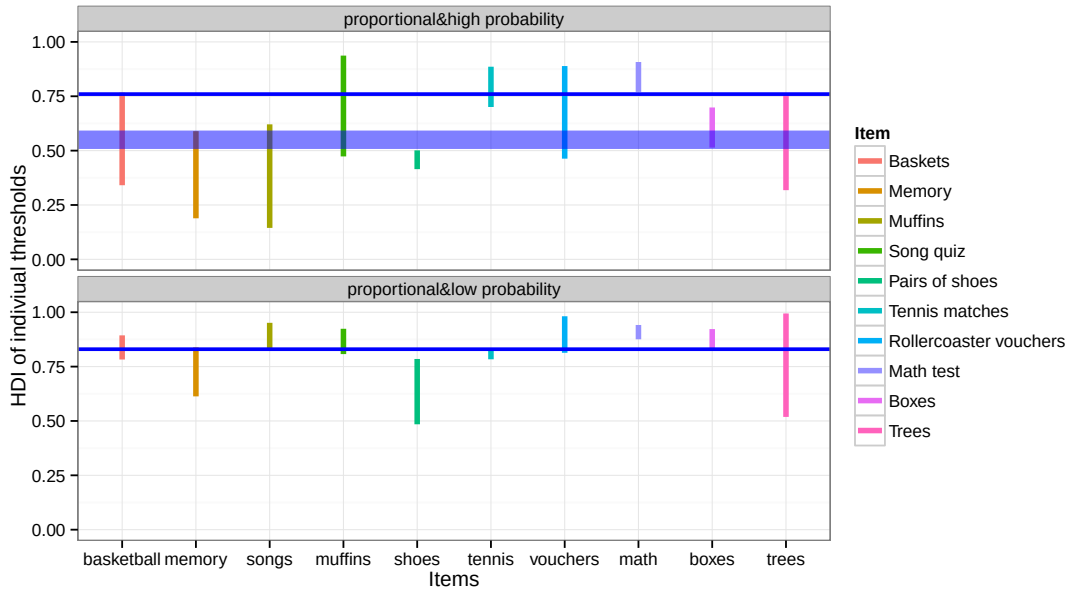
Individual θs model: In total, 17 of the 34 thresholds' HDIs overlapped in the GTM's posterior mean for θ_{many}, which was 0.83. Most of these contexts contained proportional *many*. 12 out of the 20 proportional contexts' HDIs overlapped in 0.83, among them 8 in the low probability condition. For cardinal *many*, 8 out of 14 items' thresholds overlapped in the interval [0.58, 0.59] and 7 out of 14 overlap in [0.66, 0.72] (see Figure 7). DIC = 1881.0 and pD = 96.4.

7.5. Model comparison

We want to test the hypothesis by choosing the model with the best trade-off between complexity and fit to the data. To decide on the best of all converging models, we apply a Bayesian model-choice method called deviance information criterion (DIC) which was introduced in Spiegelhalter et al. (2002). This is "a Bayesian analogue of classical model-choice criteria, such as the Akaike information criterion (AIC). DIC combines a measure of model fit - the expected deviance - with a measure of model complexity - the effective number of parameters" (Plummer 2008). This criterion is particularly suitable for the method we apply since it is simple to calculate using Markov chain Monte Carlo (MCMC) simulation and is already implemented in the program JAGS (Plummer 2010). The DIC is widely used in Bayesian statistics (cf. Plummer 2008). A high value of the DIC indicates a lot of deviance of the model's predictions from the data it is applied to. This is undesirable, of course. At the same time, the model should stay as concise as possible and not include unnecessary parameters. This is measured by the *pD*, the effective number of parameters, a measure of model complexity. The higher the pD, the more free parameters.



(a) HDIs of threshold values for cardinal *many*



(b) HDIs of threshold values for proportional *many* in high and low probability condition

Figure 7: Model for both cardinal and proportional *many*

data set	GTM	ITM	TRM	deviance GTM from ITM	GTM from TRM
proportional	977	889		10%	
cardinal	1076	995		8 %	
proportional & cardinal	2061	1881	2056	10%	0.002%

A comparison of the models' DIC states that, for each data set, the ITM yields the best fit to the

data. However, the difference between the DIC values of ITM and GTM is very small. For each data set the difference is not higher than 10%. For the data set of proportional and cardinal items, there is basically no difference between the GTM and TRM. Furthermore, we find that the HDIs of the individual thresholds differ slightly for each model. We could not find one θ which predicts the participants' behavior for each context.

8. Discussion

Statistically, the outcome of the model comparison leaves no doubt. Even though the individual threshold models are much more complex, their fit to the data is the better than the other two models'. Furthermore, for none of the three ITMs did the individual items' thresholds overlap in one interval. This seems to contradict our tentative suggestion of a unified semantics of *many* which successfully predicts the interpretation of *many* in every context or at least uniformly in each reading. Nevertheless, for all three data sets the difference in the DIC value between the three versions of the model was small. This encourages us not to dismiss of the idea of one context-independent parameter in the semantics of *many* too easily. Our findings suggest that there is an interaction with context.

The model comparison shows that allowing individual thresholds results in a better fit to the data. However, the models containing one or two general thresholds are not much worse in their fit and in predicting the measured interpretations. And we also have to keep in mind that choosing one model over the other in terms of fit to the data does not come without a cost. If we only focus on this factor, we might have to accept models which are extremely complex and this contradicts our understanding of language learning. Even though the individual-threshold models can explain the experimental data better, they are much more complex because they include one extra parameter per item. Our data set was very restricted so that each model only had to account for data from 10 to 34 contexts. This is not what a learner of a language who encounters vague expressions has to face. Her data set is substantially larger. If we assumed that a learner tried to figure out a threshold - and consequently a new meaning for *many* - for each of these numerous contexts, we would also assume that such a model would become increasingly and arbitrarily complex. This cannot be a reasonable prediction of how language learning works. So even if the ITM results in a slightly better fit to the data set, we have to keep in mind that this set is very restricted and that the predictions this model makes are not what we assume of the data set a learner faces in reality.

We also want to point out that the model we proposed is very basic and only takes into account the listener's behavior. Since the model does not predict production behavior, it ignores the fact that a listener reasons about why a speaker chose a certain word to express the meaning she wants to convey. Furthermore, the model is a semantic model, not a pragmatic one. It does not take into account alternative utterances and their complexity. These factors are only some examples from a large list of possibilities of how the model could be developed and extended. Another exciting option would be to apply it to other vague expressions like gradable adjectives.

Another interesting finding of the present approach is that it suggests different thresholds for proportional and cardinal uses of *many*. The threshold for cardinal *many* estimated at 0.69% of the cumulative density mass seems to be lower than for proportional *many* (0.83%). These values seem reliable because they were reproduced by the TRM. Nevertheless, the difference in terms of fit to the data between GTM and TRM was vanishingly small. Further research is needed to see whether, by looking at production data as well, a uniform threshold-hypothesis could be maintained after all.

A. Proportional *many*, interpretation study

1. **basketball** — Alex took part in a basketball competition and was allowed 9/12 shots from the three-point line. — HIGH: Alex, who is a professional player, made many (of the) shots. — LOW: Alex, who is an amateur player, made many (of the) shots. — How many (of the) shots do you think Alex made?
2. **memory** — For a memory test 9/12 three-digit numbers were read out to Chris. — HIGH: Chris, who has a great memory, memorized many (of the) numbers. — LOW: Chris, who has a bad memory, memorized many (of the) numbers. — How many (of the) numbers do you think Chris memorized?
3. **songs** — In a music quiz the beginnings of 9/12 pop songs were played. — HIGH: Heidi, who loves pop songs, recognized many (of the) songs. — LOW: Heidi, who hates pop songs, recognized many (of the) songs. — How many (of the) songs do you think Heidi recognized?
4. **muffins** — There were 9/12 muffins on the kitchen table in Ed's flat. — HIGH: Ed, who arrived feeling hungry, ate many (of the) muffins. — LOW: Ed, who arrived feeling full, ate many (of the) muffins. — How many (of the) muffins do you think Ed ate?
5. **shoes** — Melanie had to choose which among 9/12 pairs of shoes to bring on holiday. — HIGH: Melanie, who loves fashion, packed many (of the) pairs of shoes. — LOW: Melanie, who doesn't care about fashion, packed many (of the) pairs of shoes. — How many (of the) pairs of shoes do you think Melanie packed?
6. **tennis** — Bruno played 12/16 tennis matches last season. — HIGH: Bruno, who is an unathletic person, lost many (of the) matches. — LOW: Bruno, who is a fit person, lost many (of the) matches. — How many (of the) matches do you think Bruno lost?
7. **vouchers** — Carla won 9/12 vouchers for roller coaster rides on a fair. — HIGH: Carla, who is an adventurous person, used many (of the) vouchers. — LOW: Carla, who is a fearful person, used many (of the) vouchers. — How many (of the) vouchers do you think Carla used?
8. **math** — A math teacher presented a tricky problem to the 18/24 students in his course. — HIGH: Many (of the) students in his course, which focuses on problem-solving strategies, could solve the problem. — LOW: Many (of the) students in his course, which does not teach problem-solving strategies, could solve the problem. How many (of the) students do you think could solve the problem?
9. **boxes** — When moving to a new flat, Martha packed 15/20 boxes. — HIGH: Martha, who is a strong woman, carried many (of the) boxes herself. — LOW: Martha, who is a weak woman, carried many (of the) boxes herself. — How many (of the)boxes do you think Martha carried?
10. **trees** — Jim had 15/20 trees in his garden. — HIGH: Jim, who is a strong man, cut down many (of the) trees. — LOW: Jim, who is a weak man, cut down many (of the) trees. — How many (of the) trees do you think Jim cut down?

B. Cardinal *many*, prior elicitation and interpretation study

1. **book** — A friends favorite book has been published only recently (and has few/many pages). — How many pages do you think the book has? — intervals: 0-40, 41-80, 81-120, 121-160, 161-200, 201-240, 241-280, 281-320, 321-360, 361-400, 401-440, 441-480, 481-520, 521-560, 560 or more
2. **movie** — Nick is a man from the US (who saw few/many movies last year). — How many movies do you think Nick saw last year? — intervals: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-38, 39-41, 42 or more
3. **poem** — A friend wants to read you her favorite poem (which has few/many lines). — How many lines do you think the poem has? — intervals: 0-3, 4-7, 8-11, 12-15, 16-19, 20-23, 24-27, 28-31, 32-35, 36-39, 40-43, 44-47, 48-51, 52-55, 56 or more
4. **burger** — Joseph is a man from the US (who ate few/many burgers last month). — How many burgers do you think Joseph ate last month? — intervals: 0-1, 2-3, 4-5, 6-7, 8-9, 10-11, 12-13, 14-15, 16-17, 18-19, 20-21, 22-23, 24-25, 26-27, 28 or more
5. **shoes** — Melanie is a woman from the US (who owns few/many pairs of shoes). — How many pairs of shoes do you think Melanie owns? — intervals: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-38, 39-41, 42 or more
6. **bus** — Vehicle No. 102 is a school bus (which has seats for few/many passengers). — How many passengers do you think can sit in Vehicle No. 102? — intervals: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70 or more
7. **class** — Erin is a first grade student in primary school. (There are few/many children in Erins class.) — How many children do you think are in Erins class? — intervals: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-38, 39-41, 42 or more
8. **hair** — Betty is a woman from the US (who washed her hair few/many times last month). — How many times do you think Betty washed her hair last month? — intervals: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-38, 39-41, 42 or more
9. **friends** — Lelia is a woman from the US (who has few/many friends). — How many friends do you think Lelia has? — intervals: 0-1, 2-3, 4-5, 6-7, 8-9, 10-11, 12-13, 14-15, 16-17, 18-19, 20-21, 22-23, 24-25, 26-27, 28 or more
10. **cook** — Tony is a man from the US (who cooked himself few/many meals at home last month). — How many meals do you think Tony cooked himself at home last month? — intervals: 0-3, 4-7, 8-11, 12-15, 16-19, 20-23, 24-27, 28-31, 32-35, 36-39, 40-43, 44-47, 48-51, 52-55, 56 or more
11. **tshirts** — Liam is a man from the US (who has few/many T-shirts). — How many T-shirts do you think Liam has? — intervals: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-38, 39-41, 42 or more
12. **facebook** — Judith is a woman from the US (who has few/many Facebook friends). — How many Facebook friends do you think Judith has? — intervals: 0-69, 70-139, 140-209, 210-279, 280-349, 350-419, 420-489, 490-559, 560-629, 630-699, 700-769, 770-839, 840-909, 910-979, 980 or more
13. **coffee** — Andy is man from the US (who drank few/many cups of coffee last week). — How many cups of coffee do you think Andy drank last week? — intervals: 0-1, 2-3, 4-5, 6-7, 8-9, 10-11, 12-13, 14-15, 16-17, 18-19, 20-21, 22-23, 24-25, 26-27, 28 or more
14. **calls** — Lisa is a woman from the US (who made few/many phone calls last week). — How many phone calls do you think Lisa made last week? — intervals: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70 or more
15. **restaurants** — Sarah is a woman from the US (who went to few/many restaurants last year). — To how many restaurants do you think Sarah went last year? — intervals: 0-3, 4-7, 8-11, 12-15, 16-19, 20-23, 24-27, 28-31, 32-35, 36-39, 40-43, 44-47, 48-51, 52-55, 56 or more

References

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

Bates, D., R. Kliegl, S. Vasishth, and H. Baayen (2015, June). Parsimonious Mixed Models. *ArXiv e-prints*.

Bates, D., M. Maechler, B. Bolker, and S. Walker (2013). lme4: Linear mixed-effects models using eigen and s4. *R package version 1*(4).

Clark, H. H. (1991). Words, the world, and their possibilities. In G. R. Lockhead and J. R. Pomerantz (Eds.), *The Perception of Structure: Essays in Honor of Wendell R. Garner*, pp. 263–277. American Psychological Association.

Fernando, T. and H. Kamp (1996). Expecting many. In T. Galloway and J. Spence (Eds.), *Linguistic Society of America SALT*, Ithaca, NY: Cornell University, pp. 53–68.

Hackl, M. (2000). *Comparative quantifiers*. Ph. D. thesis, MIT.

Hörmann, H. (1983). *Was tun die Wörter miteinander im Satz? oder wieviele sind einige, mehrere und ein paar?* Göttingen: Verlag für Psychologie, Dr. C.J. Hogrefe.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Partee, B. (1989). Many quantifiers. In J. Powers and K. de Jong (Eds.), *5th Eastern States Conference on Linguistics (ESCOL)*, pp. 383–402.

Plummer, M. (2008). Penalized loss functions for bayesian model comparison. *Biostatistics*, 1–17.

Plummer, M. (2010). Jags version 3.1. 0 user manual.

Schoeller, A. and M. Franke (2015). Semantic values as latent parameters: Surprising few & many. In *Semantics and Linguistic Theory*, Volume 25, pp. 143–162.

Solt, S. (2009). *The semantics of adjectives of quantity*. Ph. D. thesis, The City University of New York.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(4), 583–639.