Semantic values as latent parameters:
Testing a fixed threshold hypothesis for cardinal readings of *few* & *many*

Abstract

Certain uses of vague quantifiers *few* and *many* intuitively compare a true quantity to *a priori* expectations about that quantity. A concrete proposal for the truth conditions of such readings stipulates a contextually-stable threshold on a contextually-variable representation of *a priori* expectations (Clark 1991, Fernando and Kamp 1996). The main goal of this paper is to introduce data-driven computational modeling as a means to implement and test complex semantic theories of this kind, which may be hard to assess based on solitary introspection of meaning intuitions. Based on an empirical measure of *a priori* expectations, we use Bayesian inference to estimate likely values of the latent threshold parameters given empirical data from production and comprehension tasks. We demonstrate how posterior inference and statistical model comparison can help assess the plausibility of the fixed threshold hypothesis.

Keywords: *few*, *many*, computational modeling, experimental data, context-dependence

1 Introduction

A striking, but well-known feature of vague quantifiers *few* and *many* is their extreme contextual variability and vagueness (e.g. Hörmann 1983, Moxey and Sanford 1993). The number of Ben's siblings needed to make (1a) true is much lower than the number of points that are needed to make (1b) true. Similarly, the number of shoes Melanie needs to own for (2a) to be true is much lower than the number of watchers in (2b). Indeed, precise truth conditions seem to be impossible to determine.

(1) a. Ben has many siblings.
b. Chris' team scored many points in the last basketball match.

(2) a. Melanie owns few pairs of shoes.
b. Few people watched the Olympics this time.

It is a challenge for linguistic theory to explain how speakers and listeners successfully communicate with expressions so context-dependent and vague and how children can acquire proficiency of their use. To address this challenge, we could try to identify a *stable core meaning* of these expressions: a complex yet systematic function from contexts to precise denotations. This paper focuses on one potential candidate for such a stable core meaning which was first suggested by Clark (1991) and formally worked out by Fernando and Kamp (1996). According to this approach, a sentence of the form "Many As are B" is true if the actual cardinality $n = |\, A \cap B \,|$ exceeds a fixed threshold θ_{many} on a measure of surprise, which is derived from a contextually supplied measure of *a priori* expectations P_E about likely values of n. In simpler terms, "Many As are B" is true if the actual number of $n = |\, A \cap B \,|$ is surprisingly high. Even with a fixed and contextually-stable threshold for what counts as sufficiently surprising, whether a certain n counts as surprisingly high can still vary dramatically for numbers of siblings and points scored during a basketball match, because we may have dramatically different prior expectations P_E. Whence that context-dependence and vagueness can be possible despite a systematic, calculable and learnable stable core meaning.

While such a surprise-based semantics may seem like an appealing idea, it also raises methodological concerns. Since the precise nature of what counts as surprising is hard to assess based on solitary introspection, it becomes exceedingly hard to test the predictions of such an account. The main contribution of this paper is therefore methodological. We seek to demonstrate how data-driven computational modeling can be a helpful addition to the linguists' toolbox, exactly where solitary introspection fails and the theory under scrutiny concerns *latent parameters* that are not directly observable, like a threshold on a measure of surprise. In other words, we argue here, by means of a case study on the meaning of *many* and *few*, for the usefulness of a particular approach to theoretically inspired statistical modeling of empirical data.

Section 2 introduces the necessary theoretical background on the meaning of *few* and *many*. Section 3 motivates our approach to data-driven computational modeling and gives the concrete model to be applied here. Section 4 describes experiments aimed to elicit representations of *a priori* expectations, as well as production and comprehension of cardinal uses of *few* and *many*. Section 5 discusses Bayesian inference of latent threshold parameters and the use of model comparison to assess the plausibility of the hypothesis that a context-independent threshold governs the experimental data from our production and comprehension tasks.

2 A fixed threshold semantics for *many* and *few*

Partee (1989) famously distinguished a *cardinal* and a *proportional* reading of *few* and *many*. The cardinal reading is exemplified in (1) and (2). According to Partee, it has a meaning "like that of the cardinal 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: 1). Truth-conditions of "Few/Many A are B" under such a cardinal reading are given in (3).[1]

(3) Cardinal reading of "Few/Many As are B"

a. *Few*: $|A \cap B| \leq x_{\text{max}}$ b. *Many*: $|A \cap B| \geq x_{\text{min}}$

Proportional readings ensue when an upper-bound on $|B|$ exists, as in (4) and (5).

(4) a. Chris ate many of the 12 muffins on the table.
b. Many Germans eat bread every day.

(5) a. Few of Martha's grandchildren could afford to buy a car when turning 18.
b. Few US adults receive the recommended amount of physical activity each week.

Partee suggests that sentence (4a) is true if Chris ate a large proportion of the muffins; at least *k*, where "[w]e may think of *k* either as a fraction between 0 and 1 or as a percentage" (Partee 1989: 2). Truth-conditions of "Few/Many A are B" under a proportional reading are given in (6).

(6) Proportional reading of "Few/Many As are B"

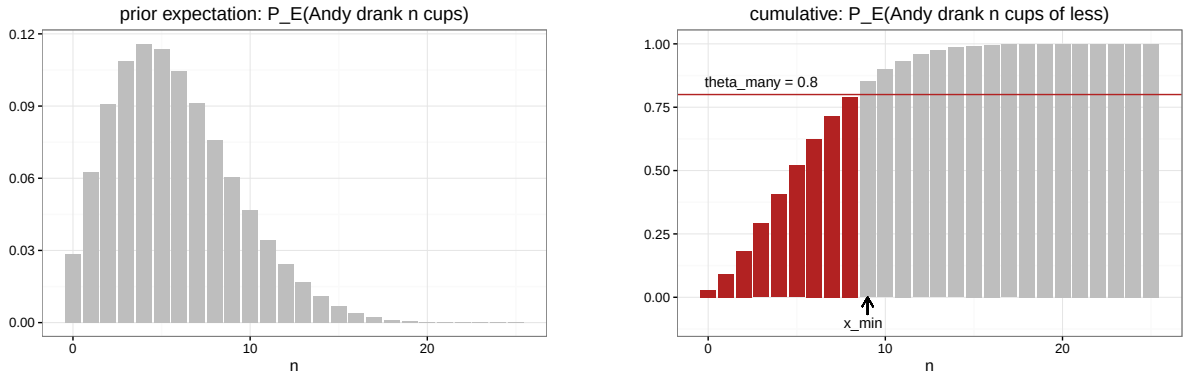
a. *Few*: $\frac{|A \cap B|}{|A|} \leq k_{\text{max}}$ b. *Many*: $\frac{|A \cap B|}{|A|} \geq k_{\text{min}}$

The semantics in (3) and (6) leave open the question of how the thresholds $x_{\text{min/max}}$ and $k_{\text{min/max}}$ are to be fixed in any given context. We will here consider one idea, which was first suggested tentatively by Clark (1991), and formally spelled out by Fernando and Kamp (1996). Let us call it the Clark-Fernando-Kamp (CFK) semantics. Neither Clark nor Fernando and Kamp commit to the idea that a single fixed threshold governs all the uses of *many* and *few*. Here, we focus on the extent to which this approach can explain in particular unstressed cardinal readings (see Section 6 for further discussion).

The idea behind the CFK semantics is that, e.g., *few* could be taken to denote "the 25th percentile (range: 10th to 40th percentile) on the distribution of items inferred possible in [the current] situation" (Clark 1991: 271). This approach explains the "cardinal surprise reading" of *few* and *many* in sentences like (7) as intensional, comparing the actual number of cups of coffee that Andy drank last week to a probabilistic belief P_E about the expected number of consumed cups of coffee in some contextually provided *comparison class* (say, American males relevantly similar to Andy, or Andy's individual coffee-drinking habits).

(7) Andy drank few / many cups of coffee last week.
$\rightsquigarrow$ Andy drank less / more cups of coffee than expected.

[1] Italicized A/B is the extension of predicate A/B.



(a) Prior

(b) Cumulative

Figure 1: Illustration of the CFK-semantics

The prior expectation P_E is highly context-dependent, both in terms of what P_E quantifies over (i.e., the comparison class) and also in terms of the precise shape of P_E. In contrast, the context-independent lexical meaning of *few* and *many* is a pair of fixed thresholds θ_{few} and θ_{many} on the cumulative distribution of P_E. Truth conditions of the CFK semantics for sentences as in (7) are given in (8).

(8) **CFK Semantics**

a. $[\![\text{Few As are B}]\!] = 1$ iff $|\, A \cap B \,| \leq x_{\text{max}}$

where $x_{\text{max}} = \max \{n \in \mathbb{N} \mid P_E(|\, A \cap B \,| \leq n) < \theta_{\text{few}}\}$

b. $[\![\text{Many As are B}]\!] = 1$ iff $|\, A \cap B \,| \geq x_{\text{min}}$

where $x_{\text{min}} = \min \{n \in \mathbb{N} \mid P_E(|\, A \cap B \,| \leq n) > \theta_{\text{many}}\}$

From (8b), the sentence "Many As are B" is true if the number $n = |\, A \cap B \,|$ is no smaller than x_{min}. In turn, x_{min} is specified as the lowest number for which the cumulative density mass of our prior expectation P_E about the number of As with property B is higher than the semantically fixed threshold θ_{many}. As a result, "Many As are B" is true if the actual number of As with property B is sufficiently surprising, where surprise is relative to contextually-variable P_E and what is sufficient surprise is encoded in contextually-stable θ_{many}.

To illustrate, consider the example in Figure 1 for the *many*-sentence in (7). Prior expectations P_E could look like in Figure 1a: they would assign a probability to any natural number n, indicating how likely we think it is that Andy drank n cups of coffee last week. Figure 1b shows the cumulative distribution of the distribution in Figure 1a. If θ_{many} was fixed to, say, 0.8, then the CFK-semantics would identify $x_{textmin}$ to be 8. Accordingly, for this P_E, the *many*-sentence in (7) would be false for any $n < 8$ and true for any $n \geq 8$.

3 Computational model

Evaluating the CFK semantics in (8) is a challenge for standard methods from theoretical linguistics insofar as they rely on intuitions about truth, entailment and the like. This is because, in almost all real-world cases, a precise enough determination of prior expectations P_E seems to elude solitary introspection. To test a CFK semantics, we therefore turn to data-driven computational modeling. For one, we use recent experimental methodology to obtain approximate empirical measures of introspectively inaccessible "prior expectations" (e.g., Kao et al. 2014, Franke et al. 2016). For another, we show how the core semantics in (8) can be turned into probabilistic models of speaker production and listener interpretation behavior. Finally, feeding empirically measured prior expectations into production and interpretation

models, we use production and interpretation data from suitable experimental tasks to infer plausible values of θ_{many} and θ_{few}.

This approach effectively considers the contextually stable thresholds θ_{many} and θ_{few} as *latent parameters*: their values cannot be directly observed but must instead be reconstructed from observable behavior. Bayesian inference is one way to do so. Given values for latent parameters, a probabilistic model makes predictions about how likely certain observable choices in production and comprehension of relevant sentences are. In technical terms, the model specifies a likelihood function $P(\text{observation} \mid \theta_{\text{many}}, \theta_{\text{few}})$ mapping values of latent parameters onto a probability of seeing a particular choice in a suitable experiment. We will use data from a production and a comprehension task to infer, via Bayes rule, which values of the latent parameters are credible, given the likelihood function and some prior over latent parameters:[2]

$$P(\theta_{\text{many}}, \theta_{\text{few}} \mid \text{observation}) \propto P(\theta_{\text{many}}, \theta_{\text{few}})\ P(\text{observation} \mid \theta_{\text{many}}, \theta_{\text{few}})\,.$$

Our goal, then, is to see whether a single pair of threshold values θ_{many} and θ_{few} explains our empirical data well enough. We focus on *many* in the exposition, but the case for *few* is parallel.

Our computational model consists of a production and a comprehension rule, both probabilistic. A probabilistic production rule is a function that assigns a probability distribution over expressions or utterances to any given meaning, while a probabilistic comprehension rule is the same in reverse, assigning a probability distribution over meanings or interpretations for each possible utterance that needs to be interpreted (e.g., Franke and Jäger 2016, Goodman and Frank 2016). Here, a production rule should give us the probability $P_S(\text{“many”} \mid n, P_E)$ with which a speaker, or speakers in general, would find the sentence “Many As are B” applicable to $n = |\,A \cap B\,|$ under prior expectation P_E. A comprehension rule should give us the probability $P_L(n \mid \text{“many”}, P_E)$ with which a listener, or listeners in general, would believe in interpretation n when they hear the relevant statement with *many* in a context where P_E captures the relevant statistical properties of the assumed comparison class.

A production rule that implements the CFK semantics in (8) is straightforward:[3] $P_S(\text{“many”} \mid n, P_E\ ;\ \theta_{\text{many}}) = 1$ if $n \geq x_{\text{min}}$ and otherwise 0, where x_{min} is derived from P_E, as in (8), based on θ_{many}, which is a free parameter for this rule (indicated by writing it after a semicolon). This probabilistic production rule is only a degenerate probabilistic rule: it only assigns the extreme values 0 and 1; it does not allow for slack, mistakes or other trembles. As such, it would not apply well to noisy empirical data. So, instead of a step-function we look at a parameterized, smoothed-out version.

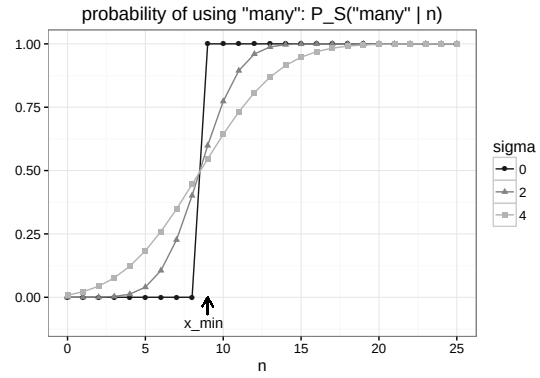
$$P_S(\text{“many”} \mid n, P_E\ ;\ \theta_{\text{many}}, \sigma) = \sum_{k=0}^{n} \int_{k-0.5}^{k+0.5} \mathcal{N}(y; x_{\text{min}}, \sigma)\mathrm{d}y \tag{1}$$

Here, σ is another free model parameter that regulates the steepness of the curve, and $\mathcal{N}(y; x_{\text{min}}, \sigma)$ is the probability density of y under a normal distribution with mean x_{min} and standard deviation σ. Essentially, this gives us a noisy implementation of speaker behavior under a CFK semantics where the amount of noise is controlled by σ. Illustrations of this probabilistic production rule are shown in Figure 2a for the example started in Figure 1. The degenerate, non-noisy production rule is the case of $\sigma = 0$.

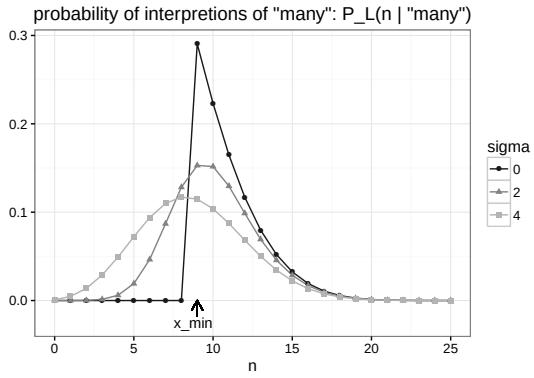
The idea behind Equation (1) is this. Assume that a hypothetically true value of θ_{many} exists. Then, given a prior expectation P_E over the contextually relevant domain, the CFK semantics in (8) gives a clear cutoff for the minimum number x_{min} of, say, cups of coffee that some particular Andy must minimally drink per week to license applicability of *many* in a sentence like (7). We should assume that speakers do not know for sure the actual x_{min} that is entailed by θ_{many} and P_E, most likely because they do not know

[2] The notation “$\propto$” for “proportional to” says that the expression on the right must yet be normalized. So, $P(x) \propto f(x)$ for some function f is short for $P(x) = \frac{f(x)}{\sum_{x'} f(x')}$.

[3] We will here propose a relatively simple computational model. For instance, we will not consider genuine pragmatic competition between alternative expressions. Other models are conceivable and may or may not give rise to similar conclusions about the tenability of a CFK semantics. We believe that this is normal: testing an abstract hypothesis (like the CFK semantics) alongside empirical data will require auxiliary assumptions about how the hypothesis relates to data observations (e.g., Quine 1951).



(a) Production rule examples



(b) Comprehension rule examples

Figure 2: Illustration of production and comphrenension rules for the example from Figure 1

P_E for certain, but that speakers nonetheless approximate it. More concretely, we assume that when a speaker decides whether some n licenses *many*, she "samples", so to speak, a noise-perturbed "subjective threshold" x'_{min} from a Gaussian distribution whose mean is x_{min} and whose standard deviation σ is a free model parameter that captures speaker uncertainty (about θ_{many}, P_E, and perhaps other things). If the sampled value is below n, the speaker finds *many* applicable to cardinality n; otherwise, she does not. This gives us a probabilistic prediction of how likely a speaker would, on occasion, find *many* applicable to n as a probabilistic function of θ_{many}, P_E and noise parameter σ.

A derivation of a reasonable probabilistic comprehension rule follows suit:

$$P_L(n \mid \text{“many”}, P_E\ ;\ \theta_{\text{many}}, \sigma) \propto P_E(n) \cdot P_S(\text{“many”} \mid n, P_E\ ;\ \theta_{\text{many}}, \sigma)\,. \tag{2}$$

This rule, which is illustrated in Figure 2b, can be motivated in two conceptually distinct ways that yield the same mathematical result. For one, we can think of Equation (2) as an application of Bayes' rule. Under this interpretation, the listener tries to infer likely world states based on a model of reverse production by taking into account how likely each world state is and how likely the speaker would use the observed *many*-statement in these states. But since the production rule in Equation (1) is just encoding "noisy truth-conditions" (rather than a genuine pragmatic choice of which out of several alternatives to use), the formulation in (2) also follows from the same considerations that motivated the production rule in (1): the formula in (2) captures interpretation based on the CFK semantics, given (Gaussian) uncertainty about threshold x_{min}.

4 Cardinal *few* and *many*

To test the CFK semantics through the lens of the computational model from the previous section we need two types of empirical data. First, we need estimates of subjects' prior expectations P_E. Second, we need data on how sentences with *few* and *many* are used and interpreted. This section presents three experiments aimed to give us such data. All three experiments use the same 14 contexts about everyday events, objects or people which all involve a quantity of some sort (see Appendix A for the full list of test items). No subject participated in more than one experiment.

4.1 Experiment 1: Prior elicitation

Design. To get an empirical estimate of participants' prior expectations, we used the *binned histogram task* of Kao et al. (2014). Participants saw descriptions of a context as in (9a) and a question as in (9b). Subjects were presented with 15 intervals, whose ranges were determined by a pre-test, and rated the likelihood that the true value lies in each interval, by adjusting a slider labeled from "extremely unlikely"

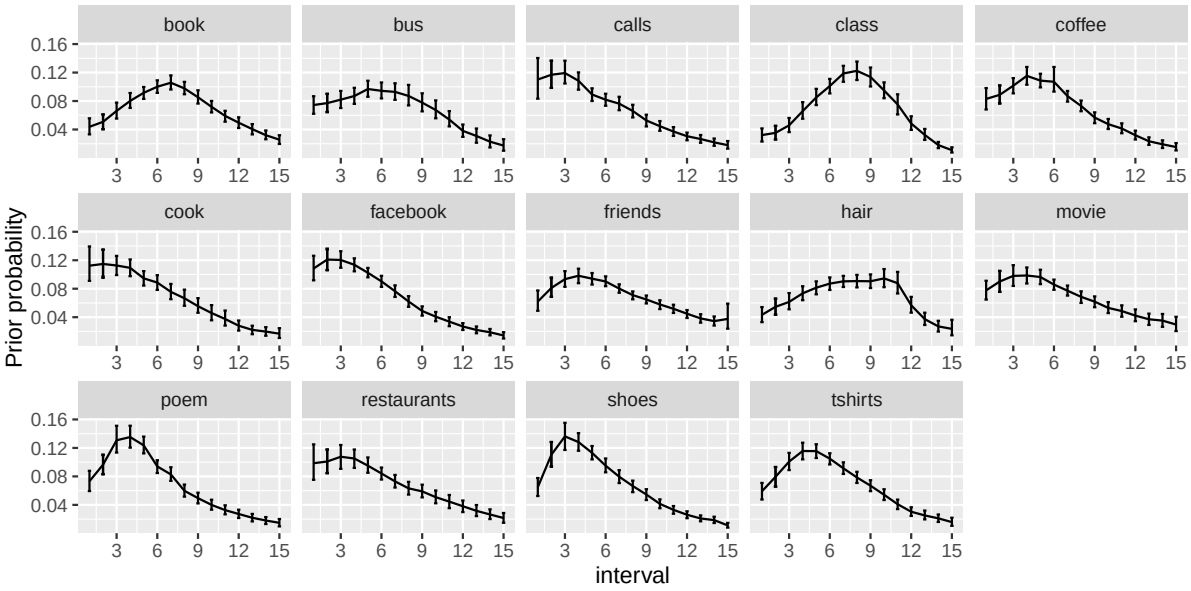


Figure 3: Empirically measured prior expectations. Error bars are estimated 95% confidence intervals.

to "extremely likely." For example, they would adjust a slider each for the probability that Andy drank 0–1, 2–3, . . . , 26–27 or more than 28 cups of coffee last week.

(9) **Prior elicitation example**

a. BACKGROUND: Andy is a man from the US.

b. QUESTION: How many cups of coffee do you think Andy drank last week?

Participants. 80 subjects were recruited via Amazon's Mechanical Turk with US-IP addresses.

Materials & Procedure. After initial instructions that explained the task, each subject saw all of the 14 contexts from Appendix A one after another. For each context, the 15 intervals were presented horizontally on the screen in ascending order from left to right. On top of each interval was a vertical slider. Participants had to adjust or at least click on each slider before being able to proceed.

Results. We excluded one participant for not being a self-reported native speaker of English. Another participant was excluded for blatantly uncooperative behavior because she had not adjusted any slider. Participants' ratings for each item were normalized and these normalized ratings were then averaged across participants. The outcome is visualized in Figure 3. These probability distributions can be conceived of as approximations to the central tendencies of the beliefs held within the population of participants (Franke et al. 2016). This average measure of P_E from Figure 3 will be input to the model.

4.2 Experiment 2: Judgment task as a production study

Design. In a binary judgment task we measured participants' production behavior of *few* and *many*. Participants were presented with a context which introduced a situation and an interval as in (10a). The interval was randomly chosen from 8 of the 15 intervals from the prior elicitation task, for example 10-12; see Appendix A. We presented only every other interval to avoid too large a number of combinations. The context was described by a statement as in (10b) which contained either *few* or *many*. Participants were asked to rate whether the statement is a good description of the context by clicking on TRUE or FALSE.

(10) **Production study example**

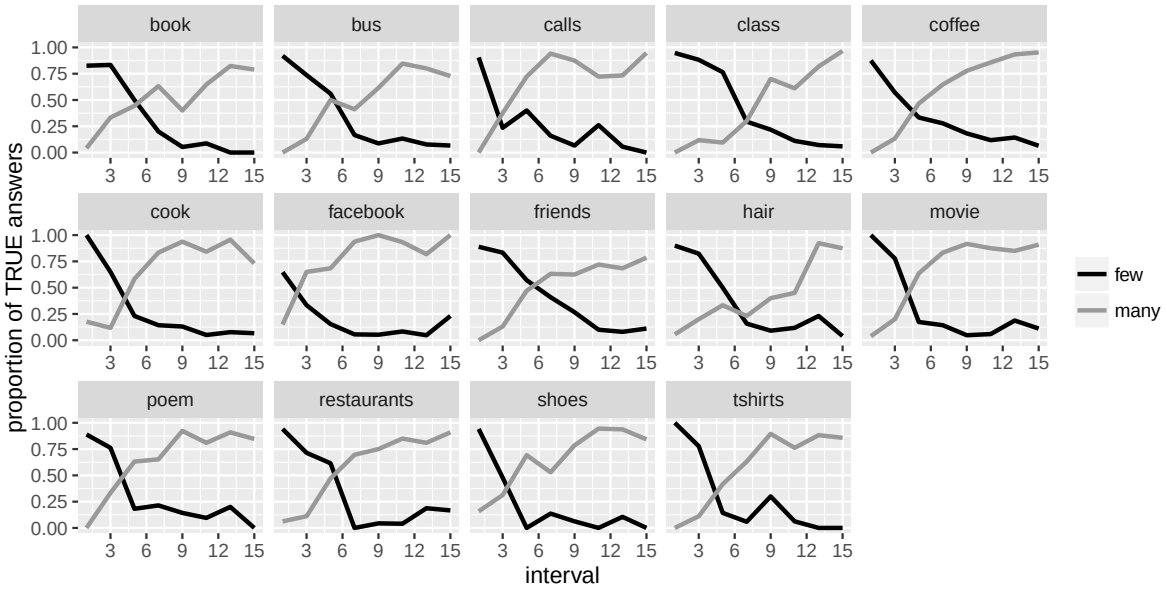


Figure 4: Proportion of TRUE answers from Experiment 2

a. CONTEXT: Andy is a man from the US who drank [2–3 | 6–7 | ... | 26–27] cups of coffee last week.
b. STATEMENT: Compared to other men from the US, Andy drank [few | many] cups of coffee.
c. QUESTION: Is this statement a good description of the context?

Participants. We recruited 301 participants with US-IP addresses via Amazon's Mechanical Turk.

Materials & Procedure. After reading a short explanation of the task, each subject saw all of the 14 contexts from Appendix A one after another. For each context, one of 8 intervals and *few* or *many* were assigned randomly. Participants had to click on one of two radio buttons labeled with TRUE or FALSE before being able to proceed to the next item.

Results. Data was excluded of 9 participants who reported not to be native speakers of English. Figure 4 shows the proportion of TRUE answers. We want the production rule P_S in Equation (1) to predict the data from this experiment.

4.3 Experiment 3: Comprehension task

Design. To measure how participants interpret *few* and *many* in different contexts, we used a forced-choice task. Participants saw descriptions of a context containing one of the quantifiers as in (11a) and a question as in (11b). They were presented with all 15 intervals for the given context and were asked to choose the interval that they thought is most likely given the background information.

(11) **Comprehension task example**
a. BACKGROUND: Andy is a man from the US who drank [few | many] cups of coffee last week.
b. QUESTION: How many cups of coffee do you think Andy drank last week?
c. 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

Participants. 200 subjects were recruited via Amazon's Mechanical Turk with US-IP addresses.

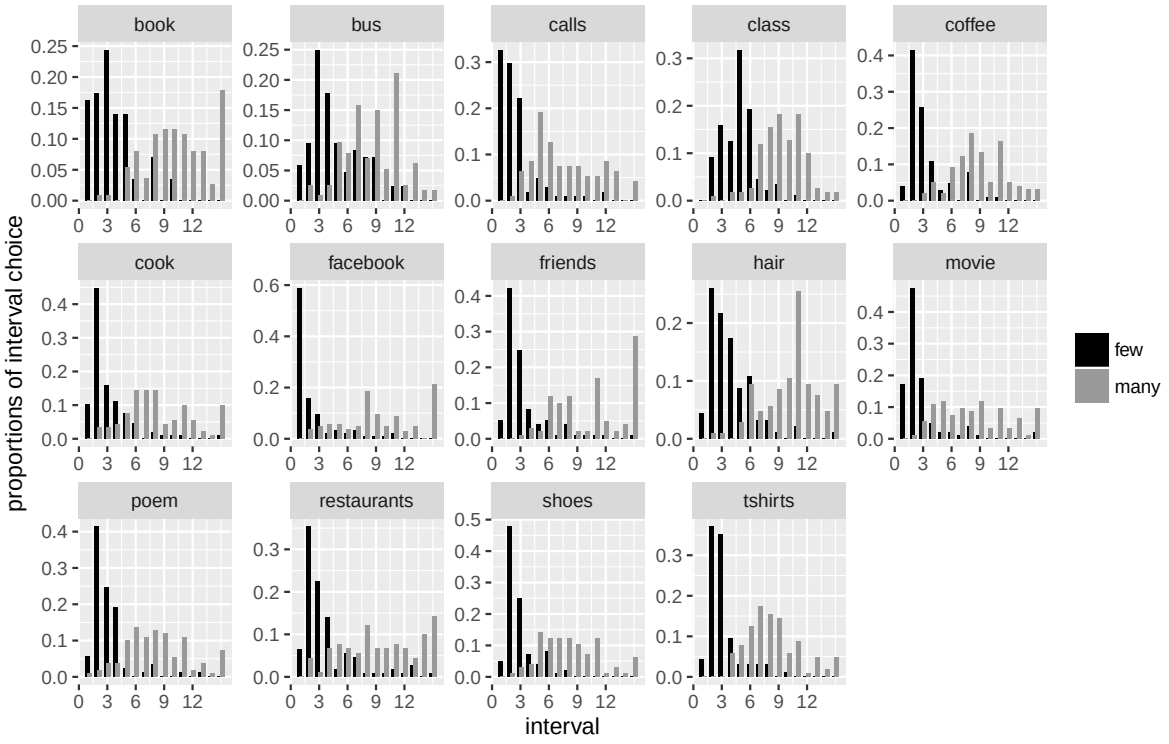


Figure 5: Proportions of interval choices from Experiment 3

Materials & Procedure. First participants read a short introduction that explained the task. Then each subject saw all of the 14 contexts in a random order. For each context, the quantifier was selected randomly and the 15 intervals were presented horizontally on the screen in ascending order from left to right. Participants had to select one interval before being able to proceed.

Results. Data from two subjects who did not identify themselves as native speakers of English was excluded. Figure 5 shows the proportions of interval choices. The comprehension rule P_L in Equation (2) is to predict the data from this experiment.

5 Model evaluation

As explained in Section 3, our goal is to learn about θ_{many} and θ_{few} from the observed experimental data. To this end, we feed the empirically measured prior expectations P_{E_i} for each item i (see Figure 3) into the production and comprehension rules in (1) and (2). This gives us likelihood functions for the production and comprehension data as described presently. We only explicitly cover the case of *many* wherever that for *few* is analogous.

Let $O_{ij}^{p_m}$ be the number of *true* answers for item i and interval j in production experiments for *many* and let $O_{ij}^{c_m}$ be the number of times interval j has been selected as the interpretation for the relevant *many*-statement about item i in comprehension experiments. Let $N_{ij}^{p_m}$ be the number of participants that saw a production trial for *many*, item i and interval j. Likewise, $N_i^{c_m}$ is the number of participants that saw a comprehension trial for *many* and item i. $O_{ij}^{p_f}$, $O_{ij}^{c_f}$, $N_{ij}^{p_f}$ and $N_i^{c_f}$ hold the same information for conditions involving *few*. Finally, let I_{ij} be the j^{th} interval of numeric values for item i. Let $\mid I_{ij} \mid$ be the length of interval I_{ij}. The probabilistic rules from Section 3 then give us (parameterized) likelihood

functions for observable data.

$$P(O_{ij}^{p_m} \mid \theta_{\text{many}_i}, \sigma_i) = \text{Binomial}\left(O_{ij}^{p_m},\, N_{ij}^{p_m},\, \sum_{n \in I_{ij}} \frac{P_S(\text{“many”} \mid n, P_{E_i}\,;\, \theta_{\text{many}_i}, \sigma_i)}{\mid I_{ij} \mid}\right)$$

$$P(O_{ij}^{c_m} \mid \theta_{\text{many}_i}, \sigma_i) = \text{Binomial}\left(O_{ij}^{c_m},\, N_{i}^{c_m},\, \sum_{n \in I_{ij}} P_L(n \mid \text{“many”}, P_{E_i}\,;\, \theta_{\text{many}_i}, \sigma_i)\right)$$

Here, $\text{Binomial}(k, n, p)$ is the probability of observing k instances of a coin coming up heads out of n coin tosses when each toss has an (independent) chance p of coming up heads.

Using Bayes rule, we can therefore make inferences about credible parameter values given the data that we observed.

$$P(\theta_{\text{many}_i}, \theta_{\text{few}_i}, \sigma_i \mid O^{p_m}, O^{c_m}, O^{p_f}, O^{c_f}) \propto \; P(\theta_{\text{many}_i}, \theta_{\text{few}_i}, \sigma_i) \cdot \prod_j P(O_{ij}^{p_m} \mid \theta_{\text{many}_i}, \sigma_i) \cdot P(O_{ji}^{c_m} \mid \theta_{\text{many}_i}, \sigma_i) \cdot P(O_{ij}^{p_f} \mid \theta_{\text{few}_i}, \sigma_i) \cdot P(O_{ji}^{c_f} \mid \theta_{\text{few}_i}, \sigma_i) \tag{3}$$

Two remarks. Firstly, we assume here that each item has its own σ_i, but that σ_i is the same for production and comprehension, as well as for *many* and *few*. This is because we think of σ_i (and the vagueness it brings) as mainly affected by uncertainty about the contextual distribution P_{E_i}. Secondly, the formula above contains as a factor the joint prior probability $P(\theta_{\text{many}_i}, \theta_{\text{few}_i}, \sigma_i)$ of parameter values θ_{many_i}, θ_{few_i} and σ_i for each item i. Here, we simply assume that θ_{many_i}, θ_{few_i} and σ_i are independent of each other and that they have uniform priors over a large-enough interval of a priori plausible values.

$$P(\theta_{\text{many}_i}, \theta_{\text{few}_i}, \sigma_i) = \text{Uniform}_{[0;1]}(\theta_{\text{many}_i}) \cdot \text{Uniform}_{[0;1]}(\theta_{\text{few}_i}) \cdot \text{Uniform}_{[0;10]}(\sigma_i)$$

To approximate the joint posterior distribution defined in (3), we used MCMC sampling, as implemented in JAGS (Plummer 2003). We collected 10,000 samples from 2 MCMC chains after a burn-in of 10,000. This ensured convergence, as measured by $\hat{R}$ (Gelman and Rubin 1992). Figure 6 shows the estimated 95% credible intervals for the marginalized posteriors over θ_{many_i} and θ_{few_i} for all items.[4]

If for all i the credible intervals for θ_{many_i} in Figure 6 overlapped, and likewise for θ_{few_i}, then this would very clearly speak in favor of a CFK semantics. Unfortunately, such clear evidence is not forthcoming. For *many*, 13 of the 14 items' credible intervals overlap in [0.687 , 0.699]. For *few*, 12 of the 14 items' credible intervals overlap in [0.148 , 0.151]. This is close to uniformity, but there are exceptions: "movies watched per year" for *many* as well as "students in class" and "facebook friends" for *few*. In effect, we do not see clear evidence in favor of a uniform CFK semantics, but we also do not see clear evidence against it.

Another possibility of assessing the idea of a uniform CFK semantics is to compare different models. The approach in (3) assumes that each item i has its own semantic threshold values θ_{many_i} and θ_{few_i}. Let us call it the Individual Threshold Model (ITM). We can compare the ITM with the outcome of a model that allows for only one θ_{many} and one θ_{few}, call this the General Threshold Model (GTM). Its posterior is defined as follows:

$$P(\theta_{\text{many}}, \theta_{\text{few}}, \sigma_i \mid O^{p_m}, O^{c_m}, O^{p_f}, O^{c_f}) \propto \; P(\theta_{\text{many}}, \theta_{\text{few}}, \sigma_i) \cdot \prod_j P(O_{ij}^{p_m} \mid \theta_{\text{many}}, \sigma_i) \cdot P(O_{ji}^{c_m} \mid \theta_{\text{many}}, \sigma_i) \cdot P(O_{ij}^{p_f} \mid \theta_{\text{few}}, \sigma_i) \cdot P(O_{ji}^{c_f} \mid \theta_{\text{few}}, \sigma_i)\,.$$

It is also possible to use information from either only the production or the comprehension data to make inferences about latent thresholds. We will make use of that possibility too in order to see whether

[4] A 95% credible interval is, intuitively put, an interval of values that are sufficiently plausible to warrant belief in (see Kruschke 2014). For example, a 95% credible interval for θ_{many_i} of $[0.6; 0.8]$ for some item i would tell us that, given the data used to condition the inference, we should be reasonably certain that the true value of θ_{many_i} is in $[0.6; 0.8]$.

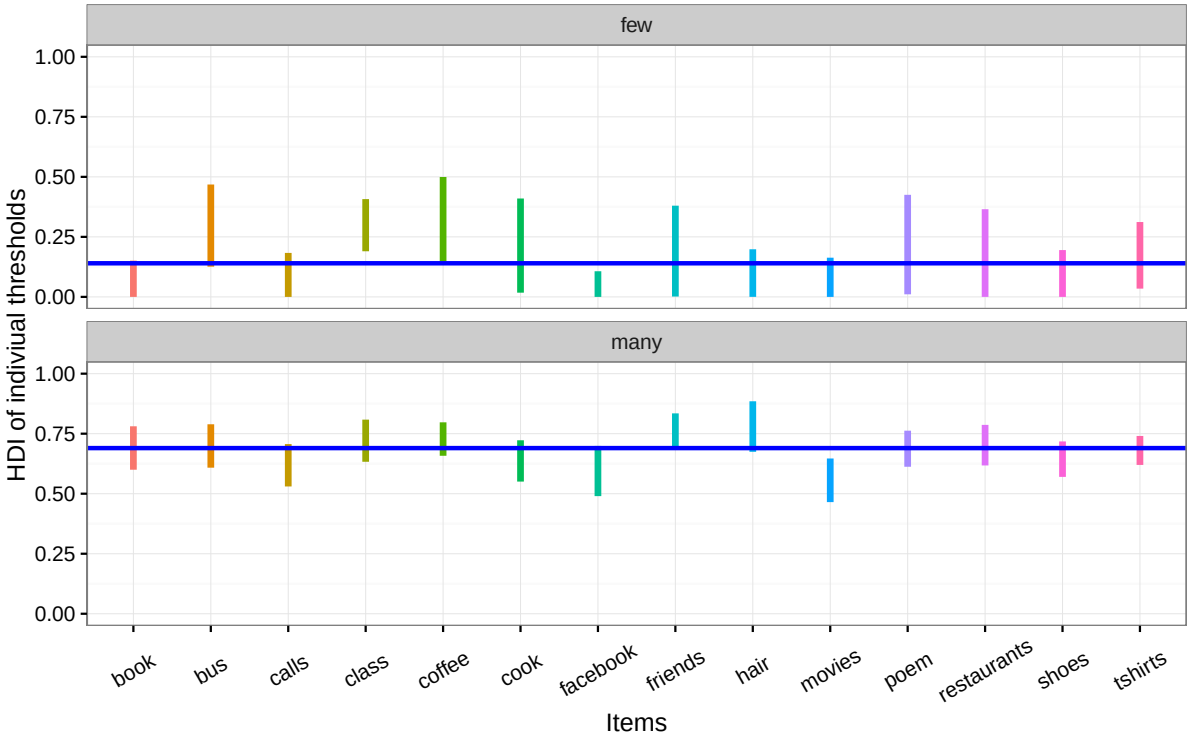


Figure 6: 95% HDIs of the estimated posteriors for thresholds for different contexts i. The horizontal lines give the biggest interval in which most contexts' HDIs overlap.

a uniform CFK semantics might work well for production or comprehension only. For example, an inference about likely item-specific thresholds based on production data only would use the posterior distribution given by:

$$P(\theta_{\text{many}_i}, \theta_{\text{few}_i}, \sigma_i \mid O^{p_m}, O^{p_f}) \propto \; P(\theta_{\text{many}}, \theta_{\text{few}}, \sigma_i) \; \cdot \prod_j P(O^{p_m}_{ij} \mid \theta_{\text{many}}, \sigma_i) \cdot P(O^{p_f}_{ij} \mid \theta_{\text{few}}, \sigma_i) \, .$$

The question we are interested in is then: which model is better suited to explain the data? This question can be addressed by statistical model comparison. There are different measures for model comparison, all based on different purposes and reasons for preferring one model over another (Vehtari and Ojanen 2012). Given our modest theoretical purposes here, we use an approach that is easy to compute based on the output of our MCMC sampling results, the so-called *deviance information criterion* (DIC) (Spiegelhalter et al. 2002, Plummer 2008). The DIC may be conceived of as a Bayesian cousin of classical model-choice criteria, in particular Akaike's information criterion (AIC). Like the AIC, the DIC weighs goodness of fit (here: the likelihood of the data given the model "trained" on the data) against the model's complexity (here: the number of its effective free parameters). Where the AIC looks at a maximum likelihood fit for the model's free parameters, the DIC consider the full posterior distribution over these, given the data. 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 number of effective free parameters, a measure of model complexity. Higher values of pD suggest higher model complexity.

Table 1 gives estimated DICs for the GTM and the ITM, based only on production data, based only on comprehension data and based on both data sets at once. We see that the GTM is roughly equal to, if not better than the ITM based on the production data only. It is a bit worse based on interpretation data and both data sets combined. Still, both models are clearly in the same ballpark. What the GTM misses in terms of goodness of fit, it makes up in terms of reduced model complexity. Based on our data alone,

	data used		
model	production	interpretation	both
GTM	DIC $= 4191.6$, $pD = 16.0$	DIC $= 2239.6$, $pD = 17.0$	DIC $= 6546.7$, $pD = 16.5$
ITM	DIC $= 4196.0$, $pD = 37.9$	DIC $= 2182.4$, $pD = 46.1$	DIC $= 6529.5$, $pD = 40.2$

Table 1: Estimated DIC values and effective free parameters

there is no clear reason to prefer either model in terms of DICs. That means that there is no reason, provided by our data, to reject the "null assumption" that a single θ_{many} and a single θ_{few} governs the use of *many* and *few*. The alternative model ITM did not do any better.

What is more, the ITM allows no possibility to generalize beyond the 14 items used here. Put differently, the ITM would assume that θ_{many} would be anywhere between 0 and 1 (its prior) for a context which was not part of the data used to condition it on. The GTM would be able to use its posterior distribution for θ_{many}. The utter lack of generalizability in ITM speaks, at least conceptually, in favor of GTM. Whether this is an empirical advantage would have to be tested. Given the data at hand and the fact that the ITM is not obviously better for this data set, there is no good reason to dismiss the hypothesis that a single pair of fixed thresholds θ_{many} and θ_{few} may have generated the production and interpretation data that we have seen.

6 Discussion and Conclusion

This paper tried to make a methodological contribution, exemplifying a potential use of data-driven computational modeling in formal semantics/pragmatics. By measuring subjects' prior expectations about real-world events experimentally, we set out to test a proposal for a semantics of *few* and *many* that is hard to assess introspectively. We showed how to couch the CFK semantics for *few* and *many* in a probabilistic model for production and comprehension. With the help of this model, we inferred *a posteriori* credible values for latent threshold parameters θ_{many} and θ_{few} from experimental data that aimed to measure production and comprehension behavior. Posterior credible values of individual threshold parameters θ_{many_i} and θ_{few_i} for different experimental items i are very similar, with overlap in the 95% HDIs of almost all items. Moreover, statistical model comparison in terms of DICs does not favor a model with individual thresholds for each item over a more parsimonious model that assumes only one fixed threshold for *many* and one for *few*. The question whether a fixed threshold CFK semantics is plausible can be answered positively, at least for the data set at hand.

The benefits of theoretically informed statistical modeling of this kind are many. The computational model makes explicit all modeling assumption including any link hypotheses regarding how theoretical notions relate to each other in producing the observable data (e.g. Chemla and Singh 2014, Franke 2016). The model considered here, for instance, assumes that the production and comprehension data are only driven by considerations of truth. In other words, the model assumes that participants in, say, Experiment 3 would not reason about what other expressions a speaker may have used other than *many* or other than *few*. This is a stark simplification. The benefit of probabilistic modeling is not only in bringing these assumptions and simplification to the fore, but in providing direct means of testing whether they are correct or, by means of model comparison, which link hypotheses may actually be better suited to explain the data.

The methodological approach introduced here opens a number of interesting venues for future research. Firstly, inference of latent thresholds could naturally be applied beyond our example case of *few* and *many*. Context-dependent threshold values are also assumed to form part of the semantics of gradable adjectives (Kennedy and McNally 2005, Kennedy 2007) and of other vague quantifiers like *most* (Hackl 2009). Computational models in combination with experimental data put themselves forward as a promising method to investigate these phenomena within a uniform framework.

Secondly, we can use probabilistic modeling to compare the CFK semantics against alternatives. For

example, a different account for the meaning of *few* and *many* was proposed by Solt (2011). Here, the threshold is derived as a positive or negative deviation from the median of the comparison class. This theory can just as well be couched in a probabilistic model and its predictions can then be compared against the CFK semantics, using statistical model comparison.

Thirdly, as mentioned briefly in Section 2, it is an open issue whether a CFK semantics, as formulated here, can also account for other readings of *many* and *few*. Fernando and Kamp (1996) apply a similar idea also to proportional readings. But there may be even more potential readings of *few* and *many*, such as the *inverse proportional reading* of (12) that would make the sentences true if the proportion of Scandinavians among Nobel prize winners was bigger than the proportion of people from other contextually salient alternative world regions who won a Nobel prize (c.f. Westerståhl 1985, Eckardt 1999, Cohen 2001, Romero 2015).

(12) Many SCANDINAVIANS won the Nobel prize.

(13) Inverse proportional reading of "Few/Many As are B"

a. *Few*: $\frac{|A \cap B|}{|A|} \leq \frac{|\bigcup Alt(A) \cap B|}{|\bigcup Alt(A)|}$ b. *Many*: $\frac{|A \cap B|}{|A|} \geq \frac{|\bigcup Alt(A) \cap B|}{|\bigcup Alt(A)|}$

It could be hypothesize that it is just be a matter of specifying the right P_E to account for these cases as well within a CFK-approach. For the inverse proportional reading of (12) in (13) we would need to consult the cumulative probability of the actual number of Scandinavians with a Nobel prize to an expectation P_E that takes, presumably, the average number of Nobel laureates in the set of all relevant world regions. It would need to be seen how far the CFK-approach can be pushed in this direction (c.f. Fernando and Kamp 1996). Still, data-driven computational modeling seems like just the right tool to help in this investigation.

Finally, it would be interesting to not only infer plausible threshold values but to try to *explain why* we see the threshold values that we apparently see. Focusing on the case of gradable adjectives, Lassiter and Goodman (2015) give a model that suggests that threshold values are the result of pragmatic inferences; another approach tries to explain why particular threshold values as evolutionarily optimal for successful communication (Franke 2012, Qing and Franke 2014). Testing these theoretical accounts with data-driven inferences of credible thresholds and applying statistical model comparison would be a natural next step.

A Experimental material

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. **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
3. **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
4. **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
5. **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
6. **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
7. **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

8. **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
9. **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
10. **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
11. **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
12. **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
13. **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
14. **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

References

Chemla, E. and R. Singh (2014). Remarks on the experimental turn in the study of scalar implicature (part i & ii). *Language and Linguistics Compass 8*(9), 373–386, 387–399.

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.

Cohen, A. (2001). Relative readings of *Many*, *Often*, and generics. *Natural Language Semantics 9*, 41–67.

Eckardt, R. (1999). Focus and nominal quantifiers. In P. Bosch and R. van der Sand (Eds.), *Focus*, pp. 166–187. Cambridge: Cambridge University Press.

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.

Franke, M. (2012). On scales, salience & referential language use. In M. Aloni, F. Roelofsen, and K. Schulz (Eds.), *Amsterdam Colloquium 2011*, Lecture Notes in Computer Science, Berlin, Heidelberg, pp. 311–320. Springer.

Franke, M. (2016). Task types, link functions & probabilistic modeling in experimental pragmatics. In F. Salfner and U. Sauerland (Eds.), *Proceedings of Trends in Experimental Pragmatics*, pp. 56–63.

Franke, M., F. Dablander, A. Schöller, E. D. Bennett, J. Degen, M. H. Tessler, J. Kao, and N. D. Goodman (2016). What does the crowd believe? a hierarchical approach to estimating subjective beliefs from empirical data. In *Proceedings of CogSci*.

Franke, M. and G. Jäger (2016). Probabilistic pragmatics, or why bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft 3*(1), 3–44.

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 457–472.

Goodman, N. D. and M. C. Frank (2016). Pragmatic language interpretation as probabilistic inference. under review.

Hackl, M. (2009). On the grammar and processing of proportional quantifiers: *Most* versus *more than half*. *Natural Language Semantics 17*(1), 63–98.

Hörmann, H. (1983). *Was tun die Wörter miteinander im Satz?, oder, Wieviele sind einige, mehrere und ein paar?* Verlag für Psychologie.

Kao, J. T., J. Y. Wu, L. Bergen, and N. D. Goodman (2014). Nonliteral understanding of number words. *PNAS*.

Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and philosophy 30*(1), 1–45.

Kennedy, C. and L. McNally (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language 81*(2), 345–381.

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

Lassiter, D. and N. D. Goodman (2015). Adjectival vagueness in a Bayesian model of interpretation. *Synthese*, 1–36.

Moxey, L. M. and A. J. Sanford (1993). *Communicating Quantities*. Hillsdale, NJ: Lawrence Erlbaum.

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. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, and A. Zeileis (Eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

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

Qing, C. and M. Franke (2014). Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. In J. Grieser, T. Snider, S. D'Antonio, and M. Wiegand (Eds.), *Linguistic Society of America SALT*, Volume 24, pp. 23–41. elanguage.net.

Quine, W. V. O. (1951). Two dogmas of empiricism. *The Philosophical Review 60*, 20–43.

Romero, M. (2015). The conservativity of many. In *Proceedings of the 20th Amsterdam Colloquium*, pp. 20–29.

Solt, S. (2011). Vagueness in quantity: Two case studies from a linguistic perspective. *Understanding Vagueness. Logical, Philosophical and Linguistic Perspectives, College Publications*, 157–174.

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.

Vehtari, A. and J. Ojanen (2012). A survey of bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys 6*, 142–228.

Westerståhl, D. (1985). Logical constants in quantifier languages. *Linguistics and Philosophy 8*(4), 387–413.