Proceedings of SALT 25: 000–000, 2015

Semantic values as latent parameters: surprising few & many*

Anthea Schöller
University of Tübingen

Michael Franke
University of Tübingen

Abstract Based on a concrete proposal for the semantics of vague quantifiers *few* and *many* suggests unspecified parameters which are hard to assess by introspection, we argue for the potential value of data-oriented computational modeling. We demonstrate how semantic values can be estimated from experimental data and a probabilistic model of language use.

Keywords: quantifiers, computational modeling, experimental data, context-dependence

1 Introduction

A man described as having “many children” is probably thought of as having four to seven kids, whereas a basketball team in the NBA that scored "many points” is much more likely to have scored 100 points or more. The same variability in use and interpretation can be found in the word *few* as well. It is a challenge to linguistic theory to explain how speakers and listeners successfully communicate with quantifiers such as *few* and *many* even though their meaning is so vague and context-dependent. To explain this, it is desirable to maintain that there exists a stable core meaning of these words, perhaps as a complex function that takes contextual parameters as input. The case of *few* and *many* is particularly interesting, because their hypothetical contextually-stable meaning escapes even trained introspection.

In this paper, we focus on the methodological problems entailed in testing a concrete lexical semantics for *few* and *many* and advocate use of computational models and empirical data. We introduce the semantic background of *few* and *many* in Section 2, elaborate on our goal and methods in Section 3, explain how we can turn a semantic theory into a computational model of language use in Section 4 and present the experiments we conducted to test the target semantics/computational model in Section 5. Section 6 applies our model to the experimental data, before Section 7 concludes with a methodological reflection.

* We thank Fabian Dablander for practical assistance and audiences in Tübingen, Stanford and Berlin 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).

2 Semantic Background

Partee (1988) argued that *few* and *many* can be read in two ways:

(1) a. Ben had many girlfriends before he got married. **cardinal**

b. Melanie owns few pairs of shoes. **cardinal**

c. Chris ate many (of the) 12 muffins on the table. **proportional**

d. Few (of the) US citizens went to the polls in the last elections. **prop.**

Partee (1988) suggests that the quantifiers' cardinal reading (1a,b) 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 1988: 1)[1]. For (1a) this means that the number of girlfriends Ben had is large, whereas the number of pairs of shoes that Linda owns (1b) is small. An interpretation of "Few/Many A are B" under a cardinal reading is given in (2):

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

For the proportional reading, on the other hand, sentence (1c) is true if Chris ate a large proportion of the muffins; at least k. "We may think of k either as a fraction between 0 and 1 or as a percentage" (Partee 1988: 2). For *few*, sentence (1d) is true if a small proportion of US citizens went to the polls, at least k. An interpretation of "Few/Many A are B" under a proportional reading is given in (3):

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

The semantics in (2) and (3) leave open how thresholds $x_{\min/\max}$ and $k_{\min/\max}$ are to be fixed in any given context. Another interesting question is whether the same procedure of fixing $x_{\min/\max}$ for cardinal readings could also apply to fixing $k_{\min/\max}$ for proportional readings. We will leave much of these issues unaddressed here, but, to make a start at least, focus our attention on cardinal readings and the question how to fix $x_{\min/\max}$ for them.

A particular proposal for fixing $x_{\min/\max}$ in cardinal readings assumes that thresholds derive from comparisons with prior expectations, giving us what we could call "cardinal surprise readings" as in (4). (It is another open question, which we will return to in Section 7, whether all cardinal readings are cardinal surprise readings in this sense.)

1 Partee (1988) labels both variables with n. For consistency with the theory proposed in section 2 we use $x_{\max}$ and $x_{\min}$ instead.

(4) Joe eats few / many burgers.
⇝ Joe eats less / more burgers than expected (for someone from the relevant comparison class).

An intuitive semantics for (4) was first suggested tentatively by Clark (1991). Clark, citing Hörmann (1983), argues that it is impossible to provide a dictionary account for *few* and *many*. A *dictionary theory* assumes that the meaning of a word can be listed as a "a brief, partial description of some aspect of the world" (Clark 1991: 264). For the meanings of *few* and *many*, Clark argues, it is impossible to come up with a short or even a finite list of denotations, since conditions of use and interpretations vary highly between different situations. As an alternative to a dictionary account, Clark suggests that, e.g., *few* could rather 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 idea was formally spelled out by Fernando & Kamp (1996). We will call it the Clark-Fernando-Kamp (CFK) semantics. It explains the target reading of *few* and *many* in (4) as intensional, comparing the actual number of burgers that Joe eats (say, per month) to a probabilistic belief P about the expected number of consumed burgers in some contextually provided comparison class (say, American males of his age and lifestyle). While the prior expectation P is highly context-dependent, the context-independent lexical meaning of *few* and *many* is a fixed threshold on the cumulative distribution of P. We will label these thresholds θ_{few} and θ_{many}. Truth conditions of the CFK semantics for sentences as in (4) are then:

(5) **CFK Semantics**

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

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

In words, given (5b), the sentence "Many A are B" is true if the number $n = |A \cap B|$ of A that are also B, 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 over A that are also B is higher than the semantically fixed threshold θ_{many}.

To illustrate this with an example, have a look at the left side of Figure 1. Prior expectation P assigns a probability to each possible $n = |A \cap B|$. For example, this could represent how many burgers we expect an American man to eat per month. Let's focus on the top row of Figure 1. Here, the context makes us expect that about 20 burgers are consumed. We take the context-independent lexical meaning of *many*

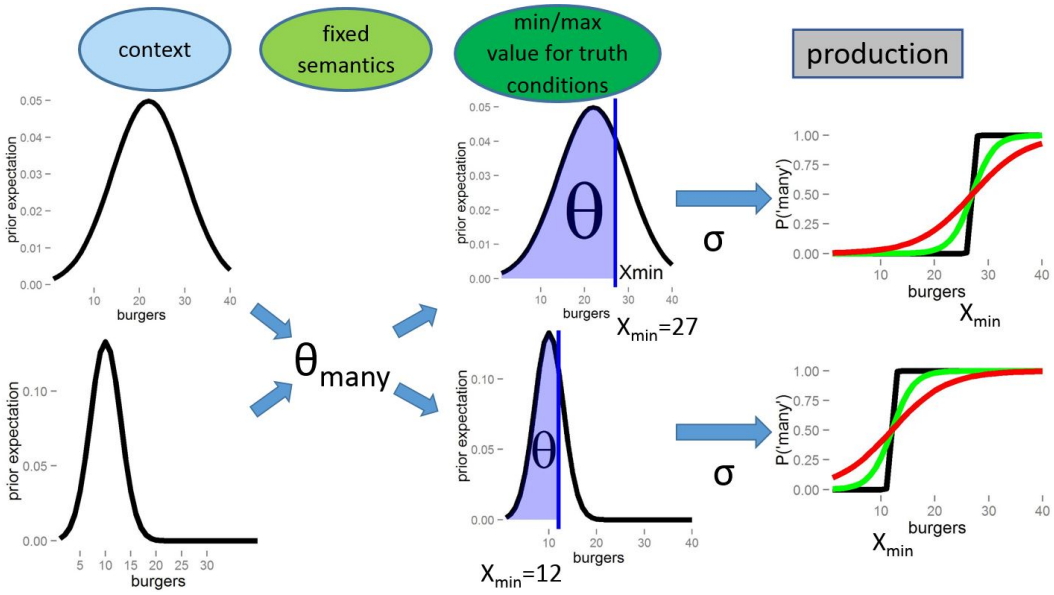

Figure 1: Applying *many*'s threshold θ_{many} to the prior expectation P determines x_{min}, fixes truth conditions. Upper and lower row show how two different contextually given P can lead to different x_{min} under a constant θ_{many}. The right hand side of the diagram shows how production predictions for *many* are derived from x_{min} (to be explained in Section 4). We use continuous functions and smooth curves for ease of presentation here.

to be a fixed threshold on the cumulative density mass of P - the area under the curve colored in blue in the graph in the middle of Figure 1. In other words, θ_{many} takes the prior expectation P of the respective context as input and cuts off the cumulative density mass of P when it has reached a percentage which is fixed in the semantics of *many*. From this cut-off we can derive x_{min}, the lowest number that is higher than the cut-off. For the context presented in the top row, "Joe eats many burgers" is true iff Joe eats more than 27 burgers. If this sentence is uttered in a different context, we will have a different prior expectation like the one in the bottom row of Figure 1, but the threshold function does exactly the same. In this case, the sentence would be true iff Joe eats more than 12 burgers.

In sum, the CFK semantics in (5) aims to explain the contextually variable thresholds x_{min} and x_{max} from the truth-conditions in (2) as a function of prior expectations P and a pair of fixed thresholds θ_{many} and θ_{few} on the cumulative distribution derived from P. Thresholds θ_{many} and θ_{few} 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.

3 Goal & Method

To assess whether the CFK semantics in (5) is on the right track is a challenge to classical methods from theoretical linguistics insofar as they rely on *intuitions* about truth, entailment and the like. This is because, in almost all cases, a precise enough determination of prior expectations about P seems to elude solitary introspection. Still, it could be the case that (5) captures speakers' non-introspective use of *many* and *few* well enough. What can we do? Certainly, we can probe intuitions (be it our own, or those of informants in a controlled experiment) about applicability and interpretation of relevant sentences in laboratory conditions that provide perfect or near-perfect information about P. This approach poses practical problems that may or may not be solvable by clever design.

But there is also an alternative that is worth exploring: data-oriented computational modeling. Focusing on *few* and *many* and the CFK semantics for their cardinal surprise uses, our main goal here is to give one constructive example of how data-oriented computational modeling could be useful for formal semantic theory. For one, we show how recent experimental methodology (e.g. Kao, Wu, Bergen & Goodman 2014) can help obtain approximate empirical measures of introspectively inaccessible "prior expectations." For another, we show how the core semantics in (5) 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 show that production and interpretation data from suitable experimental tasks can be used to infer plausible values of θ_{many} and θ_{few}.

This approach effectively considers the contextually stable semantic contribution of *many* and *few* in terms of θ_{many} and θ_{few} as a *latent variable* in a computational model that generates predictions about language use. Concretely, hypothetically fixed values for θ_{many} and θ_{few} would let us derive predictions about production and comprehension of relevant sentences. How exactly these predictions are derived from fixed latent threshold values is what a computational model has to specify. We will propose a relatively simple computational model in the next section. 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). Yet, given data and a model about how latent variables generate possible observations, we can then draw inferences about the unobservable latent variables of interest. Our goal, then, is to see whether the idea that a single pair of threshold values θ_{many} and θ_{few} explains empirical data from production and comprehension within a model of how the data is generated as a function of θ_{many} and θ_{few}.

4 Computational model

Drawing conclusions from empirical data about values of latent variables in a computational model is relatively straightforward for probabilistic models in concert with a Bayesian analysis. This is the path we trod here, as well. This section introduces probabilistic production and comprehension rules. Section 6 explains how these relate to the data from our experiments introduced in Section 5. In the following, our emphasis is on ideas, not technical detail. We focus on *many* in the exposition, but the case for *few* is parallel.

In general, a probabilistic production rule is a function that assigns a probability distribution over expressions or utterances to any given meaning (possibly subject to other parameters of interest): how likely is it that speakers would use a given expression when they want to convey a particular, fixed meaning. 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.

In our present case, a production rule should give us the probability $P_S(\text{“many”} \mid n, P)$ with which a speaker, or speakers in general, would find the sentence “Many A are B” applicable, for any possible number $n = |A \cap B|$, given expectation P over the relevant comparison class. A suitable comprehension rule gives us the probability $P_L(n \mid \text{“many”}, P)$ with which a listener, or listeners in general, would choose interpretation n when they hear the relevant statement with *many* in a context where P captures the relevant statistical properties of the assumed comparison class.

A production rule that implements the CFK semantics in (5) is straightforward: $P_S(\text{“many”} \mid n, P\ ;\ \theta_{\text{many}}) = 1$ if $n \geq x_{\text{min}}$ and otherwise 0, where x_{min} is derived from P, as in (5), based on θ_{many}, which is a free parameter for this rule (indicated by writing it after a semicolon). This probabilistic production rule is (given as the black line on the right in Figure 1) 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 be plausibly applicable to noisy empirical data. So, instead of a step-function we look at a parameterized, smoothed-out version (e.g., the colored lines on the right in Figure 1):

$$P_S(\text{“many”} \mid n, P\ ;\ \theta_{\text{many}}, \sigma) = \sum_{k=0}^{n} \int_{k-0.5}^{k+0.5} \mathcal{N}(y; x_{\text{min}}, \sigma)\mathrm{d}y\,. \tag{6}$$

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, as a function of latent parameter θ_{many}.

The idea behind (6) is this. Assume that a hypothetically true value of θ_{many}

exists. Then, given a prior expectation P over the contextually relevant domain, the CFK semantics in (5) gives a clear cutoff for the minimum number x_{min} of, say, burgers that some particular Joe must minimally eat to license applicability of *many* in a sentence like (4). We should assume that speakers do not know for sure the actual x_{min} that is entailed by θ_{many} and P, most likely because they do not know P 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, and perhaps other things). If the sampled value is below n, the speaker finds *many* applicable to cardinality n; otherwise, he 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 and noise parameter σ.

A derivation of a suitable probabilistic comprehension rule follows the exact same logic. The CFK semantics in (5) translates straightforwardly into a degenerate probabilistic comprehension rule: $P_L(n \mid \text{“many”}, P\ ;\ \theta_{\text{many}}) \propto P(n) \cdot I(n \geq x_{\text{min}})$.[2,3] Under this non-noisy rule, the listener would simply update his prior belief about the number n, given by contextually specified $P(n)$, with the information that n is no smaller than x_{min}, which he learns from assumed truth of the relevant *many*-statement. If, however, the listener is equally uncertain about the truth-conditions for *many*, e.g., due to uncertainty about comparison class distribution P, we should assume a smoothed out and parameterized comprehension rule, in analogy to the production rule in (6):

$$P_L(n \mid \text{“many”}, P\ ;\ \theta_{\text{many}}, \sigma) \propto P(n) \cdot P_S(\text{“many”} \mid n, P\ ;\ \theta_{\text{many}}, \sigma)\,. \tag{7}$$

This rule can be motivated in two conceptually distinct ways that yield the same mathematical result. For one, we can think of (7) as an application of Bayes' rule. Under this interpretation, the listener tries to infer likely world states based on a model of reverse production: 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 (6) is just encoding "noisy truth-conditions" (rather than a genuine pragmatic choice of which out of several alternatives to use), the formulation in (7) also follows from the same considerations that motivated the production rule in (6): the formula in (7) captures interpretation based on the CFK semantics given (Gaussian) uncertainty about threshold x_{min}.

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 Here, $I(\cdot)$ is the *indicator function* which takes a Boolean expression and returns its truth-value as 1 or 0 in the usual way (1 for truth).

The upshot of this is important to stress: the probabilistic production and comprehension rules that we defined here encode simple production and comprehension behavior that only take into account the semantics and noise; they are not, against superficial likeness, rules for more elaborate pragmatic production and comprehension such as variously entertained in many recent contributions (e.g. Frank & Goodman 2012; Goodman & Stuhlmüller 2013). This is where other, more complex computational models could be substituted, as suggested in Section 3. In a sense, we start with what is perhaps the simplest possible computational model. Whether adding more "pragmatics" to our model of language use changes anything about the conclusions concerning the tenability of the CFK semantics must remain to be seen. But this is orthogonal to our goal of introducing data-oriented computational modeling for estimation of semantic values.

5 Behavioral Experiments

We ran three experiments on Amazon's Mechanical Turk to assess prior expectations, production and comprehension behavior. We elicited prior expectations because they are necessary input to the model. Data measuring production and comprehension of *few* and *many* will be used to infer threshold values through the lens of our probabilistic models.

We accepted only participants with IP addresses in the United States and excluded non-native speakers of English. Each task used the same 15 contexts about everyday events, objects or people which all involved a quantity of some sort. A sample item is given below (see Appendix A for the full list of test items).

(8) a. Joe is a man from the US. **prior**

b. Joe is a man from the US who eats few/many burgers. **few/many**
How many burgers do you think Joe eats per month?

5.1 Experiment 1: Prior elicitation

80 Participants saw descriptions of a context as in (8a) and a question as in (8). To measure the participants' prior expectation of the contexts, we used the methodology of Kao et al. (2014). Participants were presented with 15 slider-interval pairs (the intervals we used depended on the respective item, determined by the pre-test) and rated the likelihood of each interval range, by adjusting a slider.

Results. Participant's ratings per item were normalized by subject-item-condition and subsequently averaged over item-condition. This gave us an empirical measure for P which will be input to the model. Figure 4 shows the probability distribution of the prior expectations which we measured for the 15 items.

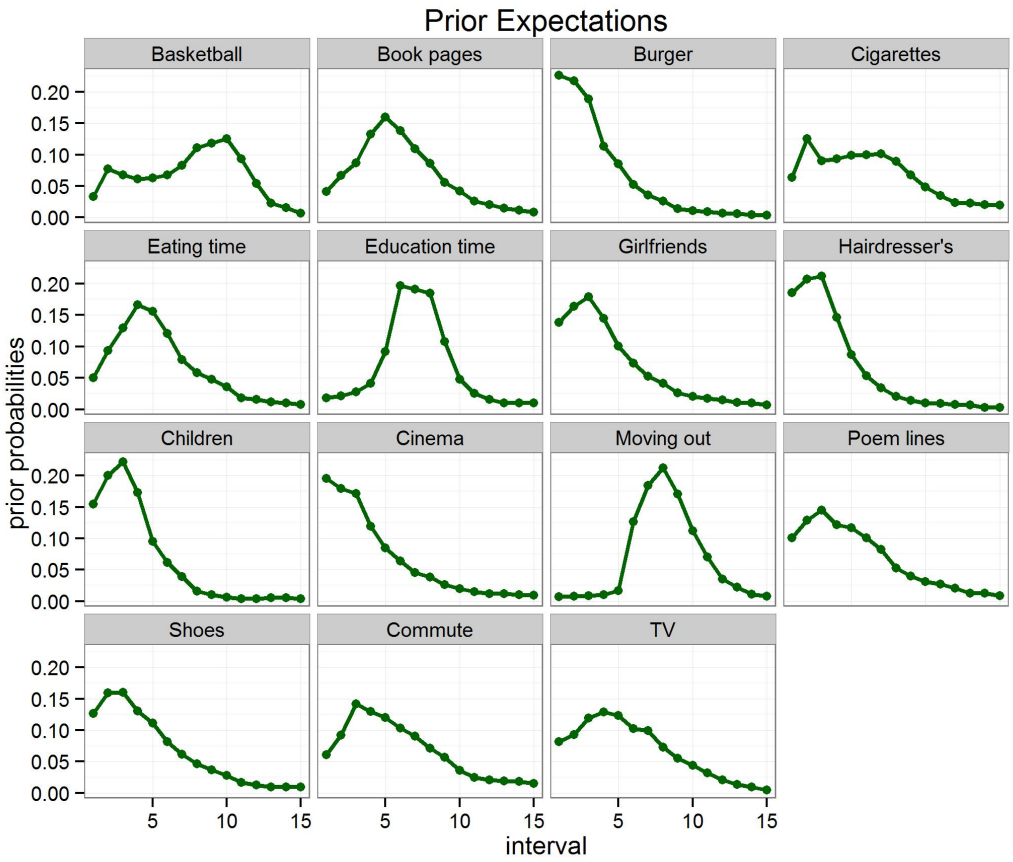

Figure 2: Prior expectation of the quantities in the 15 contexts

5.2 Experiment 2: Judgment task as a production study

We used a binary judgment task to test production of quantifiers. We asked participants how good a statement with a quantifier describes a given context. On Amazon's Mechanical Turk 350 participants read context sentences labeled as facts, which describe the context and introduced a quantity range (one of the item's intervals was randomly chosen). Participants rated whether a statement containing *few* or *many* is an adequate description of the fact. Note that in the prior elicitation task we presented participants with 15 intervals. This task presented participants with quantity ranges from only 7 of the 15 intervals to avoid too large a number of combinations. We chose the intervals with even numbers. See (9) for an example.

(9) **Fact:** Joe is a man from the US who eats **10-12** burgers a month.

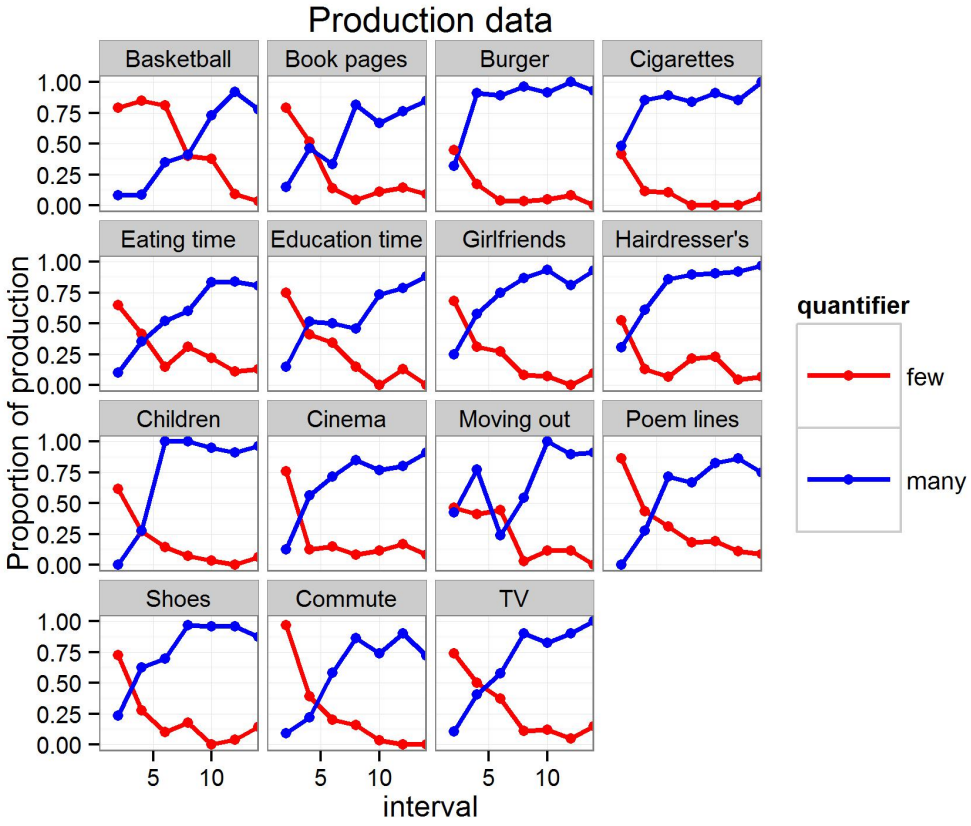

Figure 3: Production data indicating a speaker's likelihood to use a quantifier to describe an interval

Statement: Compared to other men from the US, Joe eats **few** burgers a month.
Is this statement a good description of the fact?

Results. For each item-quantifier-interval combination we looked at the percentage of yes-answers. We interpret this an indication of the likelihood that a speaker would use the quantifier to describe a quantity in the respective interval.

5.3 Experiment 3: Comprehension task

To measure how language users interpret sentences with the quantifiers we presented 60 participants with sentences introducing the contexts and *few* or *many* to describe the amount in question. See (8b) for an example. Participants were asked select the

Figure 4: Chosen numerical denotation as interpretation of the quantifiers

interval (out of all 15) they thought most likely to be the one the speaker had in mind when uttering the sentence.

Results. The histograms in Figure 4 show proportions of how often an interval was chosen as the interpretation of *few* or *many*. Count data from this experiment will be fed into our probabilistic comprehension rule.

6 Model Evaluation

The probabilistic production and comprehension rules from Section 4 implement noisy versions of the CFK semantics. They take a contextually given distribution P as input and have free parameters θ_{many}, θ_{few} and σ. 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_i, where i ranges over the 15

experimental items from Section 5, into the production and comprehension models from Section 4. This gives us likelihood functions for the data from the production and comprehension experiments as described here. (For ease of exposition, we only explicitly cover the case of *many* wherever that for *few* is analogous. The focus is again on general ideas, details about the actual implementation are left in the background.)

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 took the production experiment for *many* for item i and interval j. Likewise, $N_i^{c_m}$ be the number of participants that took the comprehension experiment 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 (as given in Appendix A). Let $\left| I_{ij} \right|$ be the length of interval I_{ij}. The probabilistic rules from Section 4 then give us (parameterized) likelihood functions for observable data:

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

$$P(O_{ij}^{c_m} \mid \theta_{\text{many}_i}, \sigma_i) = \text{Binomial}\left(O_{ij}^{c_m}, N_i^{c_m}, \sum_{n \in I_{ij}} \frac{P_L(n \mid \text{“many”}, P_i \,;\, \theta_{\text{many}_i}, \sigma_i)}{\left| I_{ij} \right|} \right).$$

Here, $\text{Binomial}(k, n, p)$ is the probability of observing k out of n coin tosses come up heads 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:

$$\begin{aligned} (10) \qquad P(\theta_{\text{many}_i}, \theta_{\text{few}_i}, \sigma_i \mid O^{p_m}, O^{c_m}, O^{p_f}, O^{c_f}) \propto \;& P(\theta_{\text{many}_i}, \theta_{\text{few}_i}, \sigma_i) \cdot \\ & \prod_j P(O_{ij}^{p_m} \mid \theta_{\text{many}_i}, \sigma_i) \cdot P(O_{ji}^{c_m} \mid \theta_{\text{many}_i}, \sigma_i) \cdot \\ & \prod_j P(O_{ij}^{p_f} \mid \theta_{\text{few}_i}, \sigma_i) \cdot P(O_{ji}^{c_f} \mid \theta_{\text{few}_i}, \sigma_i) \,. \end{aligned}$$

A few remarks about this latter formula. Firstly, we assume here that each item has its own σ_i because uncertainty about the contextual distribution P_i might be different for different items. If σ_i captures (mainly) uncertainty about P_i (like we assume here), uncertainty in production and comprehension and for *many* and *few* should be (roughly) equal. That is what (10) assumes.

Secondly, (10) also assumes that each item i has its own semantic threshold values θ_{many_i} and θ_{few_i}, contrary to the idea behind the CFK semantics that there is a

uniform, fixed threshold value pair for all items. This is because we are interested in testing whether this assumption is tenable: we would like to find out whether, given our data and our model, it is plausible to maintain that there is a fixed pair of values θ_{many_i} and θ_{few_i}: estimating individual threshold values for each item, we look for overlap in the individual estimates.

Thirdly, 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:

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

With this, we estimated posterior credible values of θ_{many_i}, θ_{few_i} and σ_i with JAGS (Plummer 2003). We collected 14,000 samples from 2 MCMC chains after a burn-in of 3,000. This ensured convergence, as measured by $\hat{R}$ (Gelman & Rubin 1992). Figure 5 shows the estimated 95% credible intervals for the marginalized posteriors over θ_{many_i} and θ_{few_i} for all items. A 95% credible interval is, intuitively put, an interval of values that are sufficiently plausible to warrant belief in (cf. 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 our data from production and comprehension and the assumption that our computational model is correct, we should be reasonably certain that the true value of θ_{many_i} is in $[0.6; 0.8]$. (In other words, credible intervals are what you would think they are; they are what confidence intervals are often mistaken for.)

What we are interested in most is to compare these credible intervals across items and to check whether we find an overlap between them. If the items' credible parameter values for θ_{many_i} overlap on some interval, this interval is where a uniform semantic threshold could be. By inspection of Figure 5, we see that there is no interval that all credible intervals for θ_{many_i} share; neither is there one that all credible intervals for θ_{few_i} share. This would suggest that, when estimating the best fitting θ_{many_i} and θ_{few_i} for each item alone, there is no single value for either threshold that is credible for *all* of our items. This would speak against the predictions of a uniform CFK semantics.

On the other hand, it is not as if credible regions for semantic thresholds appear to be completely scattered and entirely unsystematic either. If we allow for the possibility that some of our items are outliers or troublemakers for some explicable reason, a CFK semantics can still be defended. For *few* we find that the credible intervals of 13 out of 15 items overlap in [0.001, 0.006] and that credible intervals of 10 items overlap substantially on a much bigger interval. For *many*, we find overlap of credible intervals for 11 of 15 items in [0.74 , 0.76]. What this means is that our model and our data suggest that a uniform CFK semantics is tenable for at least a

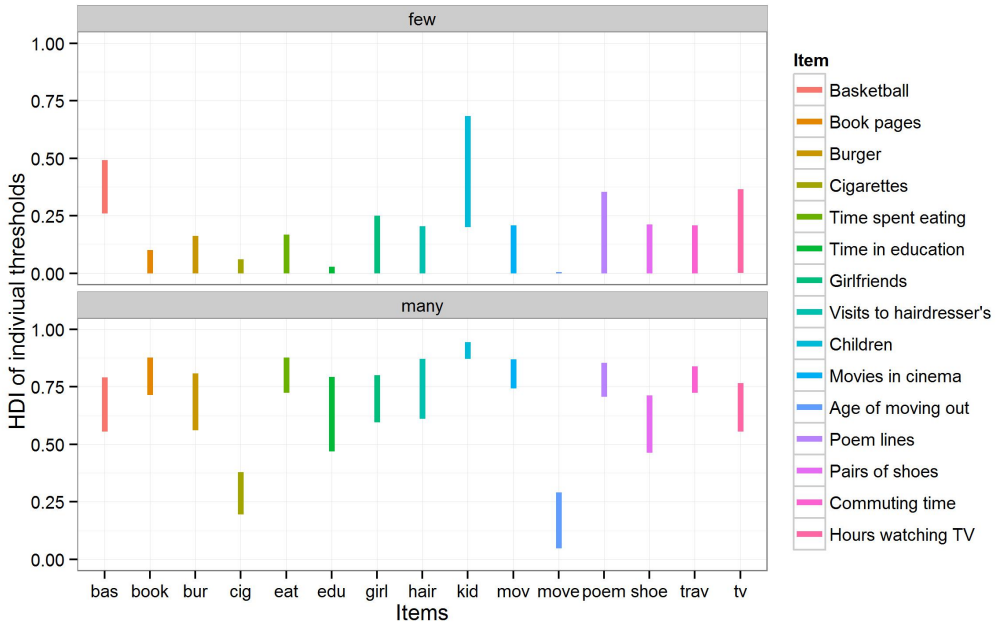

Figure 5: Credible intervals for the estimated posteriors for θ_{many_i} and θ_{few_i}.

majority of our items. We will consider possible explanations below as to why not all items seem to support a uniform CFK semantics. From the point of view that every model is wrong (including our computational model and the CFK semantics itself), but some models can be tools to shed light on interesting aspects of reality, we suggest that awareness about troublemaker cases and a reflection why they defy a uniform CFK analysis is exactly what data-oriented computational modeling should enable.

7 Discussion and conclusion

This paper tried to make a modest methodological contribution, exemplifying a potential use of data-oriented 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 semantic values from experimental data that aimed to measure production and comprehension behavior.

From the point of view of our modeling approach, the question whether the

CFK semantics is plausible can be answered with a definite "maybe." We saw that a substantial number of experimental items would lead us to believe in a range of values for θ_{many_i} and θ_{few_i} that would uniformly explain the production and comprehension behavior observed for these items. This suggests that for these items the CFK semantics makes correct predictions because we can find one common threshold θ_{few} and one θ_{many} which together with the measured prior expectations correctly predict how participants would use these quantifiers.

However, the same thresholds do not appear to be suitable for *all* of our items, see Figure 5. We want to discuss two troublesome items that are conceptually interesting. The first item ("Moving out") deals with the age at which a man moves out of his parents' house, see (11).

(11) Roger is a man from the US. He lived with his parents for few/many years before moving out.
How many years do you think Roger lived with his parents?

Figure 5 shows that the model infers thresholds that are very low in comparison with the other items for both quantifiers. This might have been due to the fact that participants were confused about the comparison class against which the evaluate the meaning of the quantifiers. Because of the phrasing of item (11), at least two readings are possible. Roger might have moved out of his parents' house as a child or when he had come of age. This assumption could be supported by the apparent variance in answers in the production task (see Figure 3) as well as in the comprehension task (see Figure 4).

These considerations point to a general problem. The CFK semantics assumes as input a specification of P as the expectations in a suitable comparison class. There can be severe uncertainty as to what the comparison class is. In real conversations, comparison classes and the relevant distribution P would have to be inferred by interpreters, alongside inferences about the quantity in question. This is not part of our simple modeling approach, and points to extensions of our computational model that include *joint-inferences* about multiple contextual unknowns (e.g. Bergen, Levy & Goodman 2014; Kao et al. 2014). Such approaches would explain how statements with vague *few* and *many* could carry information about a quantity of interest *and* about the speaker's expectations at the same time. An example of conveying information about statistical expectations with a *many* statement is this:

(12) Joe eats 42 burgers a month on average. Wow! Even for a Texan / For a self-proclaimed vegan, that's a great many burgers.

Another item whose inferred threshold values appeared incompatible with that of remaining items deals with a smoker's cigarette consumption ("Cigarettes").

(13) Margaret is a woman from the US who smokes few/many cigarettes a day.
How many cigarettes do you think Margaret smokes a day?

For this item too, the model inferred threshold values that were very low, compared to the prior expectations measured. Most people judged a sentence with *few* true only for the lowest presented interval and true for a sentence with *many* for all of the other intervals (see Figure 3). We see a very similar pattern in the interpretation data in Figure 4. Maybe participants did not use the prior expectations as they did for the other items. Since smoking has fallen in disrepute in the US, people might not only use their plain "statistical" prior expectations when they form a judgment about "few or many cigarettes." They might factor in their *moral expectations* as well (cf. Égré & Cova 2014). In principle, a CFK semantics is compatible with this idea. The prior expectations P would not only have to be sensitive to statistical beliefs about, in this case, actual number of cigarettes smoked, but also to a deontic dimension about how many cigarettes should be smoked.

Taken together, uncertainty about how exactly to determine and measure prior expectations P for a CFK semantics make it additionally hard to evaluate whether this proposal is adequate. On the other hand, it would be exactly by more extensive data-oriented computational modeling that the issues that came up here could be addressed systematically. Models in which P is a product of several sources of information (comparison classes, kinds of contextual expectations (statistical, deontic, absolute cardinality, ...)) could be formulated and tested with more careful designs aimed to probe into the composition of P. This would enable testing whether there are several kinds of cardinal readings, other than cardinal surprise readings, and equally, whether proportional readings could similarly be accounted for as functions of suitably constructed prior expectations. Allowing for such complexity, yet still providing formally rigorous methods of accessing precise predictions, is exactly why we believe that (probabilistic) computational modeling is a worthwhile addition to the linguistic toolbox.

Beyond testing a CFK semantics for vague quantifiers *few* and *many*, the presented approach also opens further possibilities. 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 & 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.

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 & Goodman (2015) give a model that suggests that threshold values are the result of pragmatic inferences; Franke (2012) and Qing & Franke (2014) try to explain why particular threshold values are evolutionarily optimal for successful communication. Testing these theoretical accounts with data-driven inferences of credible thresholds would be a natural next step.

References

Bergen, Leon, Roger Levy & Noah D. Goodman. 2014. Pragmatic reasoning through semantic inference. Unpublished manuscript.

Clark, Herbert H. 1991. Words, the world, and their possibilities. In G. R. Lockhead & James R. Pomerantz (eds.), *The perception of structure: Essays in honor of Wendell R. Garner*, 263–277. American Psychological Association.

Égré, Paul & Florian Cova. 2014. Moral asymmetries and the semantics of *many*. To appear.

Fernando, Tim & Hans Kamp. 1996. Expecting many. In Teresa Galloway & Justin Spence (eds.), *Proceedings of SALT 6*, 53–68. Ithaca, NY: Cornell University.

Frank, Michael C. & Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science* 336(6084). 998. doi:10.1126/science.1218633.

Franke, Michael. 2012. On scales, salience & referential language use. In Maria Aloni, Floris Roelofsen & Katrin Schulz (eds.), *Amsterdam colloquium 2011* Lecture Notes in Computer Science, 311–320. Berlin, Heidelberg: Springer.

Gelman, Andrew & Donald B. Rubin. 1992. Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* 7. 457–472.

Goodman, Noah D. & Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling lanuage understanding as social cognition. *Topics in Cognitive Science* 5. 173–184. doi:10.1111/tops.12007.

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

Hörmann, Hans. 1983. *Was tun die wörter miteinander im satz?, oder, wieviele sind einige, mehrere und ein paar?* Verlag für Psychologie.

Kao, Justine T., Jean Y. Wu, Leon Bergen & Noah D. Goodman. 2014. Nonliteral understanding of number words. *PNAS* doi:10.1073/pnas.1407479111.

Kennedy, Christopher. 2007. Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy* 30. 1–45.

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

Kruschke, John. 2014. *Doing bayesian data analysis: A tutorial with r, jags, and stan*. Academic Press.

Lassiter, Daniel & Noah D Goodman. 2015. Adjectival vagueness in a bayesian model of interpretation. *Synthese* .

Partee, Barbara. 1988. Many quantifiers. In Joyce Powers & Kenneth de Jong (eds.), *Proceedings of the 5th eastern states conference on linguistics (escol)*, .

Plummer, Martyn. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In Kurt Hornik, Friedrich Leisch & Achim Zeileis (eds.), *Proceedings of the 3rd international workshop on distributed statistical computing*, .

Qing, Ciyang & Michael Franke. 2014. Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. In Jessi Grieser, Todd Snider, Sarah D'Antonio & Mia Wiegand (eds.), *Proceedings of SALT*, 23–41. elanguage.net.

Quine, Willard Van Orman. 1951. Two dogmas of empiricism. *The Philosophical Review* 60. 20–43.

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

A Experimental material

Items from the interpretation task; labels and intervals used in all three experiments.

(1) **Basketball** — Dave is an adult man from the US who went to see a game of his favorite NBA basketball team and his team scored few/many points. — How many points do you think Dave's team made?
0-10, 11-20, 21-30, 31-40, 41-50, 51-60, 61-70, 71-80, 81-90, 91-100, 101-110, 111-120, 121-130, 131-140, 141 or more

(2) **Book pages** — A friend's favorite book has been published only recently and has few/many pages. — How many pages do you think the book has?
0-50, 51-100, 101-150, 151-200, 201-250, 251-300, 301-350, 351-400, 401-450, 451-500, 501-550, 551-600, 601-650, 651-700, 701 or more

(3) **Burger** — Joseph is a man from the US who eats few/many burgers. — How many burgers do you think Joseph eats a month?
0-3, 4-6, 7-9, 10-12, 13-15, 16-18, 19-21, 22-24, 25-27, 28-30, 31-33, 34-36, 37-39, 40-42, 43 or more

(4) **Cigarettes** — Margaret is a woman from the US who smokes few/many cigarettes a day. — How many cigarettes do you think Margaret smokes a day?
0-3, 4-6, 7-9, 10-12, 13-15, 16-18, 19-21, 22-24, 25-27, 28-30, 31-33, 34-36, 37-39, 40-42, 43 or more

(5) **Eating time** — Lisa is a woman from the US who spends few/many minutes of her day eating. — How many minutes do you think Lisa spends eating every day?
0-13, 14-26, 27-39, 40-52, 53-65, 66-78, 79-91, 92-104, 105-117, 118-130, 131-143, 144-156, 157-159, 160-172, 173 or more

(6) **Education time** — Sue is a woman from the US who went to school for few/many years. — How many years do you think Sue went to school?
0-2, 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, 29 or more

(7) **Girlfriends** — Ben is an adult man from the US who had few/many girlfriends before he got married. — How many girlfriends do you think Ben had before he got married?
0-2, 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, 29 or more

(8) **Hairdresser's** — Cindy is a woman from the US who goes to the hairdresser's few/many times. — How many times a year do you think Cindy goes to the hairdresser's?
0-4, 5-8, 9-12, 13-16, 17-20, 21-24, 25-28, 29-32, 33-36, 37-40, 41-44, 45-48, 49-52, 53-56, 57 or more

(9) **Children** — John is a man from the US who has few/many children. — How many children do you think John has?
0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,14 or more

(10) **Cinema** — Sally is a woman from the US who saw few/many movies in the cinema last year. — How many movies do you think Sally saw in the cinema last year?
0-3, 4-6, 7-9, 10-12, 13-15, 16-18, 19-21, 22-24, 25-27, 28-30, 31-33, 34-36, 37-39, 40-42, 43 or more

(11) **Moving out** — Roger is a man from the US. He lived with his parents for few/many years before moving out. — How many years do you think Roger lived with his parents?
0-3, 4-6, 7-9, 10-12, 13-15, 16-18, 19-21, 22-24, 25-27, 28-30, 31-33, 34-36, 37-39, 40-42, 43 or more

(12) **Poem lines** — A friend wants to read her favorite poem to you which has few/many lines. — How many lines do you think the poem has?

0-5, 6-10, 11-15, 16-20, 21-25, 26-30, 31-35, 36-40, 41-45, 46-50, 51-55, 56-60, 61-65, 66-70, 71 or more

(13) **Shoes** — Linda is a woman from the US who owns few/many pairs of shoes. — How many pairs of shoes do you think Linda owns?
0-4, 5-8, 9-12, 13-16, 17-20, 21-24, 25-28, 29-32, 33-36, 37-40, 41-44, 45-48, 49-52, 53-56, 57 or more

(14) **Commute** — Tom is an adult man from the US who spends few/many minutes a day traveling to work. — How many minutes a day do you think Tom spends traveling to work?
0-6, 7-13, 14-20, 21-27, 28-34, 35-41, 42-48, 49-55, 56-62, 63-69, 70-76, 77-83, 84-90, 91-97, 98 or more

(15) **TV** — Frank is an adult man from the US who spends few/many hours a week watching TV. — How many hours a week do you think Frank spends watching TV?
0-3, 4-6, 7-9, 10-12, 13-15, 16-18, 19-21, 22-24, 25-27, 28-30, 31-33, 34-36, 37-39, 40-42, 43 or more

Anthea Schöller
Seminar für Sprachwissenschaft
Wilhelmstrasse 19
72074 Tübingen, Germany
anthea.schoeller@uni-tuebingen.de

Michael Franke
Seminar für Sprachwissenschaft
Wilhelmstrasse 19
72074 Tübingen, Germany
mchfranke@gmail.com