few and *many*: A data-driven computational model to test a simple semantic idea

It is widely known that the concrete numerical denotation of vague quantifiers *few* and *many* can vary greatly between contexts and domains of application. Still, it is desirable on theoretical grounds to maintain that there is a stable core meaning of these words, perhaps as a complex function that takes contextual parameters as input. The case of *few/many* is particularly interesting, because their hypothetical contextually-stable meaning escapes even trained introspection. Here, we focus on the methodological problems entailed in testing a concrete lexical semantics for *few/many*, and advocate use of computational models and empirical data.

CFK-semantics. For concreteness, we focus on the "cardinal surprise usage" of *few* and *many* (Partee, 1988; Fernando and Kamp, 1996) exemplified in (1).

(1) Joe eats few / many burgers. $\rightsquigarrow$ Joe eats less / more burgers than expected.

An intuitive semantics for (1) was first suggested tentatively by Clark (1991) and formally spelled out by Fernando and Kamp (1996). According to this Clark-Fernando-Kamp (CFK) semantics, the target reading of *few* and *many* in (1) is intensional and compares the actual number of burgers that Joe eats to a probabilistic belief P about the expected number of consumed burgers in the relevant comparison class. 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, similar to degree semantics for gradable adjectives (Solt, 2011):

(2) $P(\{w \mid \text{Joe eats less burgers in } w \text{ than in the actual world}\}) < \theta_{\text{few}} \ / \ > \theta_{\text{many}}$

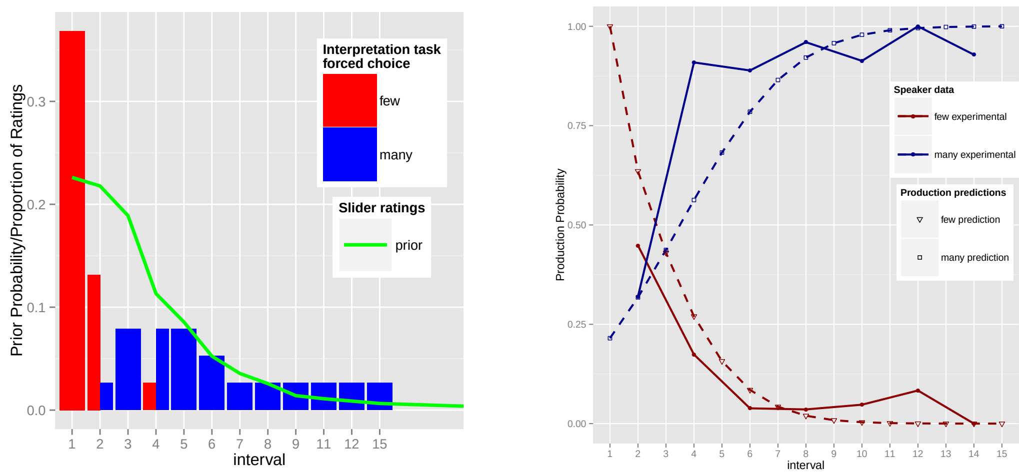
Goal. The CFK-semantics are appealing, but confront us with a methodological dilemma. How can they be tested, given that intuitions about prior expectations are hard to introspect, let alone crisp intuitions about thresholds based on them? Could it still be that our relevant conventions of use are well described by (2) on average? How could we know?

Here, we give one constructive example of how these challenges can be overcome by data-oriented computational modeling. For one, we show how recent experimental methodology (e.g. Kao et al., 2014) can help obtain reliable empirical measures of intuitively inaccessible "prior expectations." For another, we show how the core semantics (2) 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 empirical tasks can be explained well, thus corroborating the idea that (2) is a plausible generalization of the relevant use of *few* and *many*.

Model. We emphasize the crucial ideas behind our modeling and avoid technical detail. We focus on *many* in the exposition, but the case for *few* is parallel.

Let's 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 (2) 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 (1). We assume that speakers do not know for sure the actual x^*_{min} that is entailed by θ^*_{many} and P, but that speakers nonetheless approximate it. More concretely, we assume that when a speaker decides whether some x licenses *many*, he "samples", so to speak, a noise-perturbed "subjective threshold" x^s_{min} from a Gaussian distribution whose mean is x^*_{min} and whose standard deviation is a free model parameter that captures speaker uncertainty (about θ^*_{many}, P, and perhaps other things). If the sampled value is below x, the speaker finds *many* applicable to cardinality x; otherwise, he does not. This gives us a probabilistic prediction of how likely a speaker would, on occasion, find *many* applicable to x as a probabilistic function of θ^*_{many}, P and noise parameter σ (see right figure below for example predictions).

Interpretation behavior is modeled simply as reversed production, by Bayes' rule, given the probabilistic speaker behavior just sketched and the (empirically measured) prior P over x.



(a) Interpretation & prior data for "burgers" (b) Corresponding speaker production rule

Experiments. We ran three experiments on MTurk to assess priors, production and interpretation. Each used the same 15 contexts about everyday events, objects or people.

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

Prior Elicitation. 80 Participants saw descriptions like (3a) and the question in (3) and were presented with 15 intervals over relevant values x (the size of the intervals depended on the respective item, determined in a pre-test). They rated the likelihood of each interval range, by adjusting a continuous slider-bar. Ratings per item were normalized by subject-item-condition and subsequently averaged over item-condition. This gave us an empirical measure for P to feed into the model (e.g. Kao et al., 2014). See left figure, green line for an example.
Interpretation Task. 60 participants saw the same 15 item-interval pairs, but this time with additional descriptions (3b/c). Participants chose the interval they thought likely to be the one the speaker had in mind when uttering the sentence. See left figure for example data.
Production Task. 350 participants rated whether statements like in (4) were adequate descriptions of a fact (one of the 15 contexts with one of 7 intervals; see right figure).

(4) **Fact:** Joe is a man from the US who eats **10-12** burgers a month.
Statement: Compared to other men from the US, Joe eats **few** burgers a month.

Model fitting & discussion. We determined the posterior distributions of parameter values (θ_{few} and θ_{many} fixed for all 15 contexts; one σ_i per context) given the model and the data from all three tasks using MCMC sampling under suitably defined priors. For space reasons we only report point estimates and their predictions. The resulting mean estimates $\theta_{\text{few}} \approx .004$ and $\theta_{\text{many}} \approx .635$ yield a good fit to the data (correlation of observation and prediction at all data points: production $r \approx .91$, $p < 0.001$; interpretation $r \approx .64$, $p < 0.001$). This shows that it is in principle possible to uniformly explain production and interpretation data for *few* and *many* with fixed values for θ_{few} and θ_{many} under a CFK-semantics, if this is supplied by an empirically measured P and embedded within a suitable pragmatic model of lexical use and interpretation. Item-by-item analyses reveal that the low level of estimated θ_{few} still entails substantial context-variability. Although our model predicts a good "global fit", there still seem to be nuances in the use of *few* and *many*, not covered by the CFK-semantics. Time permitting, we might speculate about particularly inspiring possible meaning nuances.

References. Clark: *Words, the World, and their Possibilities* (1991). Fernando & Kamp: *Expecting Many* (1996). Kao et al.: *Nonliteral understanding of number words* (2014). Partee: *Many Quantifiers* (1988). Solt: *Vagueness in Quantity* (2011).