Typical use of quantifiers: A probabilistic speaker model

Michael Franke (m.franke@uva.nl)
Institute for Logic, Language and Computation, University of Amsterdam
Science Park, Kruislaan 107, 1098 XG Amsterdam, The Netherlands

Abstract

Many natural language quantifiers are classically associated with a stringent binary semantics, and similarly categorical pragmatic enrichments. But much experimental data shows that intuitions about appropriateness of use seem to be more fuzzy and more subtle, yet highly regular nonetheless. To account for these gradient typicality judgements, I sketch a new probabilistic model of Gricean speakers that incorporates a gradient notion of utterance alternatives. Focusing on scalar quantifier *some*, the model is a proof of concept, against expressed views to the contrary, that typicality and scalar inferences can be treated on a par.

Keywords: pragmatics of natural language; Gricean reasoning; Bayesian cognitive modeling; game theory; alternatives

Introduction

Classically, the meaning of quantifiers is described in terms of clearcut binary truth conditions (e.g. Barwise & Cooper, 1981; Peters & Westerståhl, 2006). For example, the sentence schema "Some of the *A*s are *B*s" is true just in case there is at least one *A* that is also a *B*. On top of that, it is widely held that the scalar quantifier *some*, if used in the appropriate contexts, conveys a scalar implicature, roughly, that *some but not all* of the *A*s are *B*s (c.f. Grice, 1975; Levinson, 1983). Again, this is usually treated as a categorical affair: in a given context an utterance either does or does not have an implicature.

This beautiful picture, alas, appears to be too coarse-grained. A large body of psychological literature has demonstrated that subjects' use and interpretation of quantifiers is strikingly regular but also rather fuzzy in manifold ways (c.f. Hörmann, 1983; Moxey & Sanford, 1993). Two recent studies by Degen and Tanenhaus (2011, to appear) and van Tiel (2014, to appear) showed that judgements of acceptability of sentences like "Some of the *A*s are *B*s" vary systematically but smoothly with the size of the set of *A*s that are *B*s (henceforth: the target set). Whereas these sentence are atypical descriptions when the target set size is small or when it approaches its maximum (i.e., the total number of *A*s), this is not expected under the standard categorical picture (see Figure 1 and the following section for more on typicality judgements).

It is controversial what empirically measured typicality judgements reflect. Degen and Tanenhaus (D&T) view the attested gradient patterns as evidence for a probabilistic account of pragmatic interpretation. According to their favored *constraint-based approach*, multiple factors contribute to the probability with which a listener will draw a pragmatic inference. From this point of view, gradient typicality judgements result from a fuzzy pragmatic interpretation process, of which scalar implicature calculation is a part. Unfortunately, D&T do not offer a concrete model with which to corroborate this position. In contrast, van Tiel (vT) maintains that typicality judgements are crucially different in kind from scalar implicatures. This allows him to explain away experimental evidence that might otherwise speak for a grammatical view of scalar implicature (Chierchia, Fox, & Spector, 2012; Sauerland, 2012). Obviously, then, understanding what typicality judgements are is also of theoretical significance.

Taking sides with the integrated view informally expounded by D&T, I present a probabilistic production model that aims at explaining typicality judgements on a par with scalar implicatures in a Gricean spirit. The model presented here is a conservative extension of recent Bayesian models of pragmatic reasoning as social cognition (e.g. Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013). The formal additions are (i) the integration of a richer context model, borrowed from game theory, that allows for more flexibility in modeling the implicit question under discussion, i.e., what counts as relevant to a linguistic choice, and (ii) a gradient notion of salience of alternative expressions. The latter extension is the most important one: whereas previous models of pragmatic reasoning look at a single fixed set of alternative expressions that compete in production, the present model allows for weaker or stronger activation of different alternatives and shows one possibility of integrating such a gradient notion of alternativeness in a comprehensive production model.

More concretely, the model formalizes and tests the idea that typicality ratings reflect *pragmatic appropriateness*, in particular subjects' considerations as to whether the quantifier *some* is a good lexical choice in a description of the presented situation. To determine whether a description is pragmatically well-chosen, a comparison with alternative choices is needed. I submit that subjects assess various alternatives to *some* with a variable latent probability that will be estimated from the observed data. Pragmatic appropriateness of an entertained alternative is determined based on its interpretation by an imagined listener. This presupposes an implicit goal, and so I suggest that subjects rate a quantified sentence "*Q* of the *A*s are *B*s" based on how good an answer this is to the question under discussion "How many *A*s are *B*s?"

The paper is structured as follows. The next section introduces the relevant experimental approaches to typicality of *some*. After that, I introduce the Bayesian speaker model by first motivating a parameterized representation of the question under discussion and then spelling out the probabilistic production model on top of it. Subsequently, I show that the model yields a satisfactory explanation of the data from D&T's and vT's experiments. The fitted model suggests that indeed different *a priori* natural alternatives to *some* are variably salient lexical competitors. However, as elaborated in the concluding section, these results must be relativized to a

number of modeling choices, some of which should be challenged by future work.

Typicality of *some*

Independently of each other, D&T and vT probed into the typicality structure of the quantifier *some* by having participants rate the intuitive typicality of a statement like "Some of the *A*s are *B*s" in connection with pictures that differed with respect to the cardinality of the target set. Here is just a brief description of the most important details of their respective experiments (see the original papers for further details).

Participants in vT's experiment saw pictures of ten circles in total, some number of which were black, the others white. They rated the sentence "Some of the circles are black" on a 7-point Likert scale ranging from "bad" to "good" to answer the question "How well is the picture described by the sentence?" Judgements were gathered from 30 participants via Mechanical Turk, where in each session each participant judged the critical sentence for all target set sizes in random order without interruption by filler material.

In contrast, D&T first showed participants a gumball machine with 13 gumballs. After playing a sound, a variable number of these 13 gumballs was shown as dispensed and participants were asked to rate the sentence "You got some of the gumballs" on a 7-point Likert scale ranging from "unnatural" to "natural." Additionally, D&T also included an 8^{th} option of choosing "false" on a button separated from the rating scale. Still, in their analyses the data was treated as if obtained from an 8-point Likert scale with "false" answers coded as lowest. Moreover, D&T ran four versions of their experiment. For one, they elicited typicality ratings separately for sentences with and without the partitative construction *some of the*, henceforth *summa* vs. plain *some*. For another, while all experiments did include fillers, they either did or did not include numerical expressions "You got one/two/three ... of the gumballs." Taken together, there were then four experimental versions, which I will refer to as "Some NoNum", "Some Num", "Summa NoNum" and "Summa Num" in the following. There were 120 and 240 participants for the Num and NoNum experiments respectively, but the number of judgements collected differed between target set sizes (see original paper for details).

Figure 1 shows the mean normalized typicality ratings from vT's study ("Summa vT") and the "Summa NoNum" version of D&T. The general pattern is shared also with the other versions of D&T's experiments: the quantifier *some* is rather atypical in connection with empty sets, but also with smaller quantities; typicality ratings peak at slightly below half of the maximal cardinality; typicality judgements drop for larger set sizes, albeit not as steeply as towards smaller set sizes.

Context model: goals & expression alternatives

I propose to explain intuitive typicality judgements as reflections of *pragmatic appropriateness* of the given sentence as

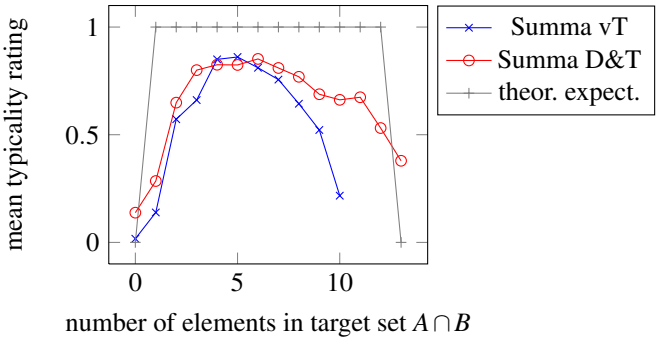

Figure 1: Mean typicality ratings of "Some of the *A*s are *B*s" for different target set sizes in different experiments (see main text). The gray line gives the expected applicability under standard linguistic theory (for $|A| = 14$).

a description of the presented picture. This presupposes an implicit goal-structure (appropriate for what?). I submit that subjects assume that the implicit question under discussion is "What is the cardinality of the target set?" with varying degrees of required precision in the answer. Judgements of pragmatic appropriateness also require a comparison to other potential descriptions.

Signaling games (Lewis, 1969) provide a rich formal model of a context of utterance that includes explicit goals and choice alternatives. A signaling game captures the interaction between an informed speaker and an uninformed listener. The speaker observes a state, but the listener does not. The speaker sends a message with a fixed conventional meaning and the listener guesses which state is actual by reasoning about the speaker's motivations for choosing the message in question over alternative options. Correct guesses are communicative successes and result in high payoffs for cooperative interlocutors in the sense of Grice (1975), while false guesses are failures and yield low payoffs. But when there are many possible world states and absolute precision is not strictly necessary, we might also allow for a gradient notion of communicative success (c.f. Jäger & van Rooij, 2007). If t is the actual state and t' is the listener's chosen interpretation, then the received payoff is a function of the similarity of t and t'. Depending on how exactly a measure of similarity is mapped onto payoffs, we can capture the idea that almost guessing the right state may be an almost perfect outcome, while farther deviations are increasingly bad.

States & utilities. Applying this idea to the stimuli of the experimental cases at hand, the states of the game are the respective cardinalities of the target set. The maximal number of black balls in vT's experiment was 10. The maximal number of gumballs in D&T's experiment was 13. Consequently, we get eleven or fourteen states $T = \{t_0, t_1, \ldots, t_{10/13}\}$ as the possible information states of the speaker. The idea that communication is successful proportional to how close the listener's guess matches the actual cardinality can be made

precise in many ways. For simplicity, I follow Jäger and van Rooij (2007) and use a simple one-parameter utility function in the vein of Nosofsky (1986) (here and below, free parameters are notationally separated by a semicolon):

$$\mathrm{U}(t_x, t_y \,;\, \pi) = \exp\left(-\pi \cdot (x-y)^2\right) .$$

The parameter π captures what level of precision is relevant: as π goes to infinity, interlocutors want to identify states precisely, but for lower values of π further deviations from the actual state will also count as sufficiently successful. For example, with $\pi = 30.25$ (roughly the mean of the estimated posteriors, see below) utilities come out as:

$$\begin{aligned} &\mathrm{U}(t_0,t_0) = 1 && \mathrm{U}(t_0,t_1) = 7.29e-14 \\ &\mathrm{U}(t_0,t_2) = 2.82e-53 && \mathrm{U}(t_0,t_3) = 5.79e-119 \\ &\mathrm{U}(t_0,t_4) = 6.33e-211 && \ldots \end{aligned}$$

Alternative expressions. Next to the interlocutors' communicative goals, another crucial ingredient of a conversational context model is a specification of alternative expressions that enter into pragmatic reasoning. Normally, game theoretic or Bayesian models assume that a handful of concrete alternative expressions compete with each other. But it seems more plausible to consider a large enough base set of potentially entertained alternative expressions, backed up by a probabilistic measure of how likely each subset of expressions is actually entertained when speakers assess the pragmatic appropriateness of a particular expression. The latent degree of salience should be estimated from the data, not hand-picked by the modeller. This is the approach I would like to take here.

As for the base set of alternative expressions to *some*, I suggest to consider its obvious scale mates *all* and *none*, but also small numerals within the subitizing range *one*, *two* and *three*, of which it is prima facie plausible that they compete with *some* when describing a scene of, say, two out of ten black circles. As for larger cardinalities, it may seem similarly reasonable to consider *most* and *many* as potential alternatives as well. Numerals are given a precise semantic interpretation, quantifier *some* its usual logical semantics, while *most* and *many* receive simple proportional semantics: *most* receives a more than ½-interpretation, while *many* receives a more than ¾-interpretation.[1]

A probabilitic production model

Following recent models of pragmatic reasoning (Benz & van Rooij, 2007; Franke, 2011; Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013; Jäger, 2013), we consider a speaker's approximately optimal language use, based on the assumption that the listener interprets messages literally. A (hypothetical) literal listener, who only acts on the semantic meaning of messages, chooses each true interpretation with equal probability:

$$P_{LL}(t \mid m) = \mathcal{U}(t \mid m \text{ is true in } t) ,$$

where $\mathcal{U}$ the uniform distribution (over state set T).

Given a belief in a literal listener, the sender's expected utility for sending m in state t is:

$$\mathrm{EU}_S(m \mid t \,;\, \pi) = \sum_{t'} P_{LL}(t' \mid m) \cdot \mathrm{U}(t, t' \,;\, \pi) .$$

A rational speaker would choose only messages that maximize expected utility. If the context affords precise communication ($\pi \to \infty$), rational choices under a belief in literal interpretation obeys Grice's Quantity Maxim of choosing the most informative expression. In this sense, we are assuming a generalization of a Gricean speaker who seeks to maximize the informativity of his utterances.

Speakers may occasionally fail to choose optimally and be maximally informative. Still, even with errors, slips and limited computational resources, speakers' choice probabilities can be expected, on average, to be proportional to their expected utility. A handy implementation of such utility-proportional choice is the *soft-max rule* (Luce, 1959; Sutton & Barto, 1998), which here takes the form:

$$P_S(m \mid t, X \,;\, \lambda, \pi) = \frac{\exp(\lambda \cdot \mathrm{EU}_S(m \mid t \,;\, \pi))}{\sum_{m' \in X} \lambda \cdot \mathrm{EU}_S(m' \mid t \,;\, \pi))} , \qquad (1)$$

where $X \subseteq M$ is the subset of alternative expressions that the speaker takes into account. The parameter λ in Equation (1) controls for the speaker's rationality in the sense that with $\lambda \to \infty$, choices adhere to the standards of pure rationality, while with $\lambda \to 0$, choices become entirely random.

Speakers may not take all conceivable alternative expressions into consideration equally. The probability of choosing a message for a given state is therefore obtained by weighing in the probability that X is entertained:

$$P_S(m \mid t \,;\, \lambda, \vec{s}) = \sum_{X \subseteq M} P(X \mid \vec{s}) \cdot P_S(m \mid t, X \,;\, \lambda, \pi) .$$

On the simplifying assumption that the salience probabilities $\vec{s} = \langle s_1, \ldots, s_{|M|} \rangle$ of entertaining individual messages are independent, the probability of entertaining a set $X \subseteq M$ of alternative expressions is:

$$P(X \mid \vec{s}) = \prod_{m_i \in X} s_i \cdot \prod_{m_i \notin X} 1 - s_i .$$

Linking function for ordinal data

The production model defined above can be used to predict, for each cardinality of the target set, how likely a speaker would use the description "Some of the *A*s are *B*s." However, the relevant data on typicality judgements is ordinal data obtained from Likert-scale rating tasks. Instead of assuming flat out that the typicality data can be treated as interval-level data (like both vT and D&T did), it is more appropriate to fit the

[1] The chosen alternatives and their semantics (especially for *most* and *many*) are clearly questionable in principle, but picked here also with practical considerations in mind. See concluding remarks.

model to the ordinal data directly. One option for doing this is to use a probit linking function, like in ordinal probit regression (see Kruschke 2011, Chapter 21).

The probit linking model entertains a vector of non-decreasing thresholds $\vec{\theta} = \langle \theta_0, \ldots, \theta_{|D|} \rangle$, with $|D|$ the number of items on the rating scale, that determine the relative sizes of the rating categories. Only $\theta_1, \ldots \theta_{|D|-1}$ are free parameters, while $\theta_0 = -\infty$ and $\theta_{|D|} = \infty$. The latent categorical choice probability $P_S(m \mid t, X\ ;\ \lambda, \pi)$ is perturbed by Gaussian noise with standard deviation σ. The probability of choosing degree d_j of the ordinal rating scale is the likelihood that the noise perturbed value falls in the interval (θ_{j-1}, θ_j). Taken together, the likelihood function is:

$$P_S(d_j \mid m, t\ ;\ \lambda, \pi, \vec{s}, \vec{\theta}, \sigma) = \sum_{X \subseteq M} P(X \mid \vec{s}) \cdot \int_{\theta_{j-1}}^{\theta_j} \mathcal{N}(x \mid P_S(m \mid t, X\ ;\ \lambda, \pi), \sigma) \mathrm{d}x\,, \tag{2}$$

where $\mathcal{N}(x \mid \mu, \sigma)$ is the probability density function of a normal distribution with mean μ and standard deviation σ.

Parameter estimation

The speaker model has 9 free parameters: λ (rationality), π (precision), and one salience probability parameter s_i for each of the seven alternatives to *some* that we chose to include. (We should naturally assume $s_{some} = 1$.) The linking function comes with additional free parameters: σ (linking noise) and $|D| - 1$ threshold parameters θ_i that determine the relative sizes and positioning of the $|D|$ ordered categories of the rating scale. Priors were chosen as follows:

$$\lambda \sim \mathcal{U}(0;30) \qquad \pi \sim \mathcal{U}(0;100) \qquad s_i \sim \mathcal{U}(0,1)$$
$$\sigma \sim \mathcal{U}(0;.4) \qquad \theta_i \sim \mathcal{N}(i/|D|, 4 \cdot |D|)\,.$$

Uniform priors were chosen in order to remain uncommitted, except for determining a credible range. The priors on the thresholds of the linking model reflect the prior assumption that the degrees would be roughly evenly spaced along the unit interval, but allow for significant deviation. Priors also encoded the ordering of thresholds.

We are interested in the joint posterior likelihoods of these parameters separately for each of the five experiments. Posteriors were estimated with MCMC sampling using JAGS (Plummer, 2003). Two chains of 5000 samples were gathered for each experiment with a thinning rate of 2 after an initial burn-in of 2500.

The most interesting part is the estimated posteriors over salience degrees s_i of competing alternatives in different experiments (see Figure 2).[2] Strikingly, if the model is correct, the data offers little support for the idea that *two*, *three*, *most* and *many* (with the semantics assumed here) are strong competitors to *some*. On the other hand, high levels of salience of alternatives *none*, *one* and *all* are likely, given the model and the data.

Furthermore, there is quite some interesting variation between experiments. Firstly, as vT's experiment provided far fewer data points, the 95% HDI intervals tend to be bigger. Secondly, vT's data is suggestive of higher levels of salience of most alternatives. This could be due to the repetitive nature of the experiment, where no fillers where included and subjects only judged the critical sentence "Some of the *A*s are *B*s." Another possibility is that *some* occurred in subject position in vT's experiments, but in object position in D&T's. Thirdly, when we compare results for D&T's experiments, we notice that the *summa* construction seems to be more strongly associated with the *all* alternative than the *some* construction. However, in the "Num"-type experiments where additional numerical expressions were interspersed between critical trials, the salience of *all* seems heavily reduced. Finally, there is further interesting variation in the estimated salience levels of alternative *one*. As expected, the presence of number expressions raised the salience of this expression. But, on top of that, the increase of salience between "Num" and "NoNum" versions seems much higher for the partitative *summa* construction. We could speculate that this might be due to the fact that the included numerical expressions were partitative constructions as well (e.g., *one of the*). However, all of the above remarks must be taken with a grain of salt, as they are not based on stringent statistical comparison and are likely to be eventually influenced by more realistic modeling (see discussion in conclusion).

Model validation

There is no other formalism that predicts typicality ratings to compare the presented model to, but to at least provide a crude measure of goodness-of-fit Figure 3 plots the means of 5000 samples from the posterior predictive distribution against the observed data. Each dot in the graph corresponds to a pair of predicted and observed numbers of choices of a degree on the relevant rating scale for each target set cardinality. Correlation coefficients between posterior predictions and observations are summarized in Table 1. From this the model's posterior predictions seem reasonable but not perfect. We should bear in mind though, that these results are

	some		*summa*		
	NoNum	Num	NoNum	Num	vT
r	.980	.988	.974	.990	.973

Table 1: Pearson's r of posterior predictions and observations.

[2] Estimates of the posteriors for other parameters, especially those of the linking function, bear no surprises and are also not of great conceptual interest. There were no noteworthy divergences in posterior credence for values of parameters λ, π and σ between experiments. Linking noise σ is estimated as slightly higher for vT's data, which is not surprising given that the scale used in this experiment had one degree less. Moreover, the estimated linking thresholds θ_i for D&T's data show a bigger gap between the first and the second threshold than any other pair of adjacent thresholds. This is likely a result of the distinct category of the lowest rating degree ("false", as opposed to "unnatural") which was presented as a separate option outside of the Likert-scale.

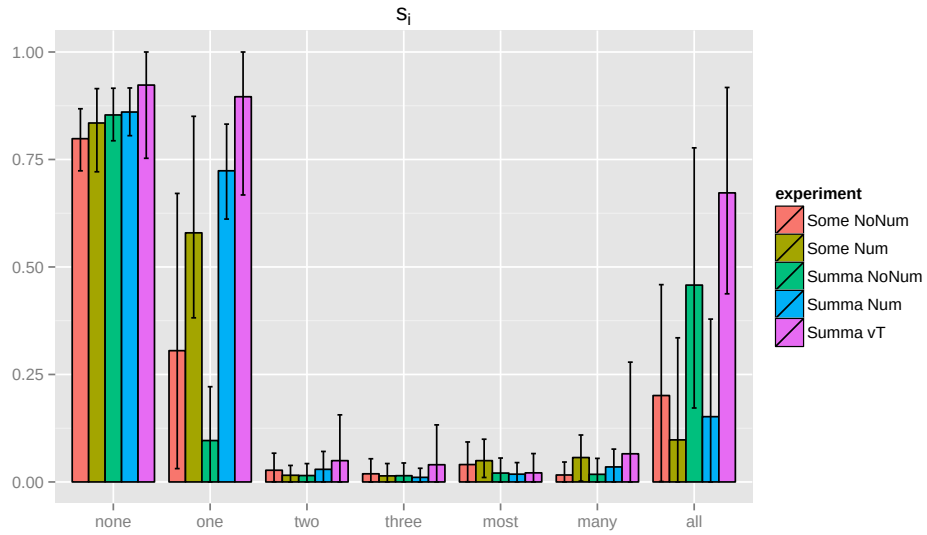

Figure 2: Posterior estimates over salience degrees of alternatives *some*. The plot shows the means of the marginalized posterior likelihoods for each alternative expression (barplot), together with 95% HDI intervals (error bars).

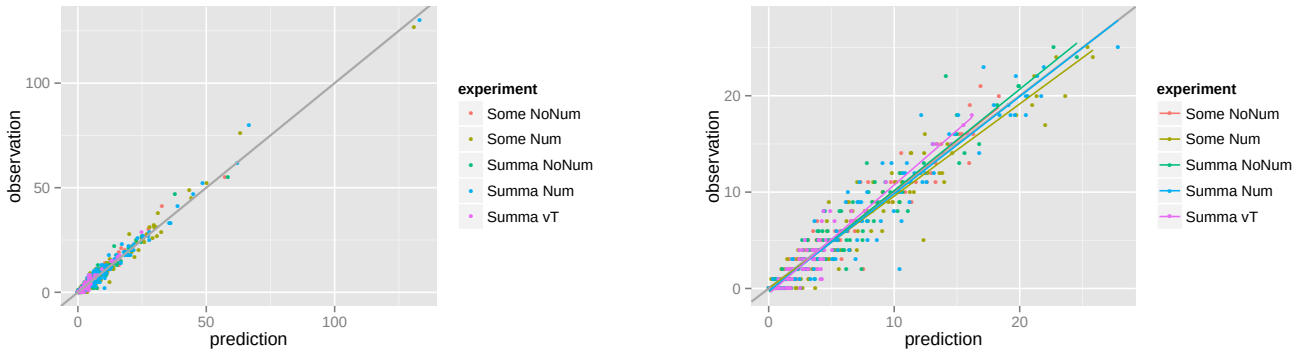

Figure 3: Prediction-observation plots. See main text for description. The gray lines are the diagonals, included for orientation. The right plot zooms in on the low number region and also includes the best linear fits (for the low-number subset).

based on the full posterior predictive distribution, not just on the maximum likelihood estimates and that we attempted to predict exact numbers of choices of each Likert-scale item.

Conclusion

The main theoretical contribution here is a proof of concept that it is possible to explain graded typicality judgements and scalar implicatures with a unified Gricean speaker model. To do so, the model extended existing approaches to pragmatic reasoning as social cognition by including (i) a pragmatic precision parameter that regulated how strictly the question under discussion ("How big is the target set?") needs to be answered; and (ii) a gradient notion of alternativeness. If the model is true, the data supports the conclusion that *none*, *one* and *all* are the most serious competitors, while *two*, *three*, *most* and *many* appear less strongly associated with *some*.

Still, there are reasons why these conclusions must be considered preliminary at best. The presented model contains a few simplifying assumptions that should be scrutinized more carefully. Lifting these assumptions may change at least the quantitative results obtained. The most striking simplifying assumptions are: (i) a small stipulated set of potential alternatives, (ii) independence of salience of alternatives, and (iii) uniform payoff structure over varying cardinalities.

Ad (i). It may appear unmotivated to fix the rather small set of potential alternatives that I considered here. What about larger numerals, or expressions like *a few*, *almost all* or *more than half*? My choice of alternatives here has been guided entirely by practical considerations. Firstly, vague expressions like *a few* and *almost all* do not have a clearcut uncontroversial semantics (the same does hold for *many*). Secondly, their inclusion would, if anything, have led to tighter predictive fit (with more uncertainty in the posteriors), because there

would have been more free parameters. The same applies to the inclusion of larger number terms. Essentially, we should not even try to justify *any* armchair choice, but to look for an empirical measure of gradient alternativeness. The modest purpose here was to provide a modeling framework in which to make sense of such future data.

Ad (ii). The model assumes that whether a speaker is aware of alternative *two* is independent of whether she is aware of *one* and *three*. This is unintuitive, but it is difficult to model potential dependencies in this sense. Eventually the modeling of salience in alternatives should be corroborated anyway by other empirical measures of lexical association.

Ad (iii). The model also assumes, again counterintuitively, that the utility structure is uniform. Similarity between t_x and t_y only depends on the differences between x and y, not on their absolute values. This might not respect the intuitive and empirically attested perceptual similarity ratings between pairs of cardinalities: low cardinalities that differ by a fixed amount appear less similar to each other, than higher cardinalities that differ by the same amount (e.g. Logan & Zbrodoff, 2003). In general, it seems clear that future work in this direction should try to combine models of pragmatic reasoning with models of (approximate) number sense and size estimation (e.g. Kaufman, Lord, Reese, & Volkmann, 1949; Atkinson, Campbell, & Francis, 1976).

Acknowledgments

I would like to thank Judith Degen and Bob van Tiel for sharing their data with me and for many helpful comments, as well as Jakub Dotlaĉil, Mike Frank, Noah Goodman, Dan Lassiter, Roger Levy and Jakub Szymanik for discussion. This work was supported by NWO-VENI grant 275-80-004.

References

Atkinson, J., Campbell, F. W., & Francis, M. R. (1976). The magic number 4 +/- 0: A new look at visual numerosity judgements. *Perception*, *5*(3), 327–334.

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

Benz, A., & van Rooij, R. (2007). Optimal assertions and what they implicate. *Topoi*, *26*, 63–78.

Chierchia, G., Fox, D., & Spector, B. (2012). Scalar implicature as a grammatical phenonenon. In C. Maienborn, K. von Heusinger, & P. Portner (Eds.), *Semantics. An international handbook of natural language meaning* (pp. 2297–2332). Berlin: de Gruyter.

Degen, J., & Tanenhaus, M. K. (2011). Making inferences: The case of scalar implicature processing. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 3299–3304).

Degen, J., & Tanenhaus, M. K. (to appear). Processing scalar implicatures: A constraint-based approach. *Cognitive Science*.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998.

Franke, M. (2011). Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics & Pragmatics*, *4*(1), 1–82.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling lanuage understanding as social cognition. *Topics in Cognitive Science*, *5*, 173–184.

Grice, P. H. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics, vol. 3, speech acts* (pp. 41–58). Academic Press.

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

Jäger, G. (2013). Rationalizable signaling. *Erkenntnis*.

Jäger, G., & van Rooij, R. (2007). Language structure: Psychological and social constraints. *Synthese*, *159*(1), 99–130.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American Journal of Psychology*, *62*(4), 498–525. Retrieved from http://www.jstor.org/stable/1418556

Kruschke, J. E. (2011). *Doing Bayesian data analysis*. Burlington, MA: Academic Press.

Levinson, S. C. (1983). *Pragmatics*. Cambridge, UK: Cambridge University Press.

Lewis, D. (1969). *Convention. a philosophical study*. Cambridge, MA: Harvard University Press.

Logan, G. D., & Zbrodoff, N. J. (2003). Subitizing and similarity: Toward a pattern-matching theory of enumeration. *Psychonomic Bulletin & Review*, *10*(3), 676–682.

Luce, D. R. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.

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

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Peters, S., & Westerståhl, D. (2006). *Quantifiers in language and logic*. Clarendon.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd international workshop on distributed statistical computing.*

Sauerland, U. (2012). The computation of scalar implicatures: Pragmatic, lexical or grammatical. *Language and Linguistics Compass*, *6*(1), 36–49.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning*. MIT Press.

van Tiel, B. (2014). *Quantity matters: Implicatures, typicality, and truth*. Unpublished doctoral dissertation, Radboud Universiteit Nijmegen.

van Tiel, B. (to appear). Embedded scalars and typicality. *Journal of Semantics*.