Words are vague: A model of generic language

Michael Henry Tessler (mhtessler@stanford.edu), **Noah D. Goodman** (ngoodman@stanford.edu)
Department of Psychology, Stanford University

Abstract

Generic utterances are ubiquitous in natural language. Despite their prevalence, the meanings of generic statements are puzzling to formal approaches. Cimpian, Brandone, and Gelman (2010) demonstrated that the endorsement rate of generic statements differs by context, and that generics can be endorsed based on weak evidence while the same sentences are interpreted strongly. Here, we replicate these effects in Exp. 1 and investigate how models based on prevalence (the probability of the property given the category) can account for this behavior. Using Bayesian data analysis techniques, we show how a simple scalar prevalence semantics is untenable, but that the same semantics within a probabilistic pragmatics framework can account for the data. In this model, the generic has an underspecified meaning, but this uncertainty is resolved by context and interaction. Context effects are predicted if the prior distribution of prevalence differs by context; in Exp. 2 we find direct evidence that this is so. We conclude by showing that the model is able to capture accidental and low-prevalence generics—two cases of theoretical importance.

Keywords: generics; pragmatics; bayesian cognition; bayesian data analysis

> New sanctions passed by this Congress, at this moment in time, will all but guarantee that diplomacy fails – alienating America from its allies; making it harder to maintain sanctions; and ensuring that Iran starts up its nuclear program again.
>
> (Barack Obama, *2015 State of the Union Address*)

Generic meanings are hard to pin down. Consider President Obama's remark during the State of the Union Address. The sentence is a generic statement about *new sanctions* in that it conveys a generalization about the members of this kind (Carlson, 1977; Leslie, 2008). It leaves totally open the question: "Exactly how many of these new sanctions will guarantee diplomatic failure?" President Obama's statement is conceivably true if most or only a few new sanctions will be problematic. At the same time, he is not a man to waste words—why does he go through the trouble of producing such a vague utterance? In this paper, we will see that the *context* in which his words are uttered is essential to the meaning we derive. We propose that this aspect of generic language follows from pragmatic reasoning about an uncertain threshold for meaning; an idea which we formalize in a probabilistic model within the Rational Speech Acts framework (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013).

Generic statements are puzzling because their meaning is so flexible. On the one hand, generics would seem to suggest an almost universal quantification, as in "Dogs bark". Others, like "Mosquitos carry West Nile virus", involve a property that applies only to a small subset of the kind. Cimpian et al. (2010) (henceforth, CBG) carried out a series of experiments designed to examine the truth conditions and implications of generic statements. They found evidence for the influence of additional knowledge about a target property (e.g. its *distinctiveness*) on participants' willingness to accept generic statements. This type of contextual information modified the truth conditions. CBG also found an asymmetry between interpretation and verification: in one task, participants interpreted a generic (e.g. "lorches have purple feathers") as nearly universal; in a different task, they would endorse the same generic as true at a much lower prevalence (e.g. when "50% of lorches have purple feathers").

Both context and asymmetry effects pose a puzzle for the semantics of generics: what could be the stable meaning of a generic given this extreme flexibility? In this paper, we seek to explain both of these phenomena as the effects of pragmatic inference filling in a meaning that is underspecified in the semantics. In particular, we posit a scalar semantics for generics in which they express that the probability of the property given the kind—–which we'll refer to as its *prevalence*—–is above a threshold (cf. Cohen (1999)). Following Lassiter and Goodman (2015), we treat this threshold as a free variable that is reasoned about by a pragmatic listener: what is the threshold likely to be, given that a speaker bothered to utter the generic? Context effects follow from differences in prior beliefs about the distribution of the property across categories. Asymmetry effects fall out of modeling task differences between the language understanding and answer-selection tasks faced by participants in the different experiments (cf. Degen and Goodman (2014)).

In what follows, we first replicate the main effects reported by CBG. We use Bayesian data analytic techniques to further examine the effective truth-conditions of generic statements. We then introduce a model of generic comprehension, within the probabilistic Rational Speech Acts framework. We show that this model predicts both context and asymmetry effects, given appropriate prevalence priors. We experimentally elicit the prevalence priors in CBG's experimental contexts, verifying the predictions of the model. We close with a short discussion and demonstration that the model can capture additional cases of theoretical importance.

Experiment 1: CBG replication

In CBG's *truth conditions* task, participants were given an evidence statement consisting of the percentage of a category that had a property (e.g. "30% of lorches have purple feathers"). Participants were asked to judge the associated generic statement (i.e. "Lorches have purple feathers") as true or false.

The authors manipulated context within-subjects by adding additional statements about the property. We focus on three contexts in this paper: *dangerous and distinct* (DD, e.g. "These feathers are as sharp as needles and can easily get lodged in you, causing massive bleeding. No other animals have these kinds of feathers"), *not distinct and irrelevant* (NI, e.g. "These feathers are wide and very smooth to the touch. Other animals have these kinds of feathers."), and *plain* (P, no additional statements). CBG found that DD increased the overall proportion of "true" responses to the generic, particularly so at lower prevalence levels.

In their *implied prevalence* task, participants were supplied with the generic (again, context was a within-subjects variable) and asked to judge prevalence: "What percentage of lorches do you think have purple feathers?". CBG found that the generic was interpreted strongly—nearly all lorches have purple feathers—in all contexts.

Experiment 1 attempted to replicate the main findings of CBG: that context affects the proportion of "true" responses to a generic statement (Exp. 1a) and that there is an asymmetry between verification and interpretation of the truth conditions of the generic (Exp. 1b). Exp. 1a and 1b were conducted on separate sessions, one week apart. None of the participants completed both experiments.

Experiment 1a: *truth conditions*

Participants We recruited 40 participants over Amazon's crowd-sourcing platform Mechanical Turk.

Procedure and materials Our procedure was very similar to CBG's *truth conditions* task. Participants were told they were the resident zoologist of a team of scientists that recently discovered an island with many new animals, and that their task was to provide their expert opinion on questions about these animals[1].

We used the same materials as CBG (available in their Appendix). The materials used were 30 novel animals (e.g. lorches, morseths, blins) each paired with a unique property. Properties were pairs of colors and body-parts (e.g. purple feathers, orange tails). Each participant saw 30 unique animal-property pairs: 10 in each of 3 contexts (*DD*, *NI*, *P*). The 10 items in each context were randomly paired with 1 of 5 "prevalence levels": $\{10, 30, 50, 70, 90\}$%; each prevalence level appeared 2 times per context.

Participants saw a prevalence statement and a context statement (*DD*, *NI*, *P*; illustrated above). Participants were then asked "Is the following sentence true or false?", below which was presented the associated generic (e.g. "Lorches have purple feathers") and "True" and "False" radio buttons.

Results Results are shown in Figure 1 (left). We entered participant's truth judgments into a mixed effects logistic regression with random by-item and by-participant effects of intercept and fixed effects of prevalence and context as well as their interaction[2]. Our results replicated the finding of CBG that the generic statements were endorsed more in the dangerous and distinctive (DD) context than in the plain (P) context (Figure 1, left; $\beta = 1.99; SE = .36; z = 5.52; p < .001$). There was also an interaction between prevalence level and context such that the generic was endorsed more in *DD* context than in *P* at lower prevalence levels ($\beta = -.03; SE = .01; z = -2.35; p = 0.019$). There was a trending effect for the *NI* context to be endorsed *less* than the plain context ($\beta = -.57; SE = .30; z = -1.91; p = .056$).

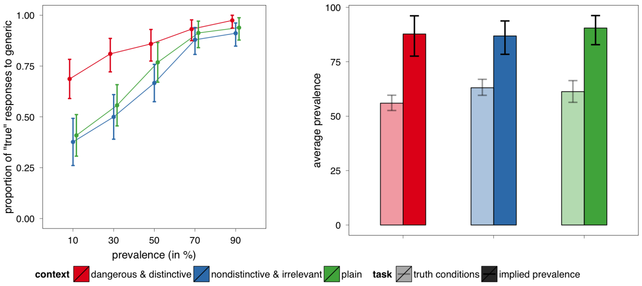


Figure 1: Replication of CBG. Left: *truth conditions* vary by context (Exp. 1a). Right: *implied prevalence* of the generic is greater than *truth conditions* (Exp. 1b).

Experiment 1b: *implied prevalence*

Participants We recruited 30 participants over Amazon's crowd-sourcing platform Mechanical Turk.

Procedure and materials Our procedure was very similar to CBG's *implied prevalence* task. Our instructions were the same as in Exp. 1a[3].

The materials and context conditions were the same as in Exp. 1a. On each trial, participants saw a generic statement (instead of a prevalence statement) and a context. Participants were then asked "What percentage of [the kind] do you think have [the property]?" (e.g. "What percentage of lorches do you think have purple feathers?") The dependent measure was a free response required to be an integer, $0 - 100$.

Data analysis and results We followed the data analysis strategy of CBG. Using the data from Exp. 1a, we computed, for each subject, an *average prevalence level* that led to "True" responses (e.g. if a participant said "True" whenever the prevalence was 70% or 90% and "False" to everything else, that participant received an *average prevalence score* of 80%). This score was compared against the implied prevalence dependent measure of Exp. 1b.

The prevalence scores from each task were entered into a linear mixed model with a by-participant random effect of intercept; the fixed effects were context, task, and their interaction. Our results replicated the asymmetry finding of CBG that the generic statement was interpreted as having a higher prevalence than its truth conditions imply (i.e. main effect of task; $\beta = 28.8; SE = 4.3; t = 6.6; p < 0.001$; see Fig. 1, right).

[1] The experiment in full can be viewed at `http://stanford.edu/~mtessler/experiments/generics/cbg2010-replication/experiment/experiment-9.html`

[2] This was the maximal mixed-effect structure supported by the data.

[3] The experiment in full can be viewed at `http://stanford.edu/~mtessler/experiments/generics/cbg2010-replication/experiment/experiment-12.html`

Fixed-threshold semantics

The above results, replicated from CBG, indirectly constrain the effective truth conditions that participants are using for generic statements within these experimental conditions. In this section we begin the model-based exploration of these truth conditions by positing a simple threshold semantics. We explore how well this could account for the data of Exp. 1 using Bayesian data analysis.

Assume that the conditions for truth of the generic can be usefully represented by a threshold on prevalence: the generic is true when the prevalence of some property within a kind exceeds a given threshold (see Cohen (1999) for a similar assumption). If we use $x \in [0,1]$ to denote the prevalence $P(\text{property}|\text{category})$, then the simple threshold meaning is:

$$g(x,\theta) = \begin{cases} 1 & \text{if } x > \theta \\ 0 & \text{if } x \leq \theta \end{cases} \tag{1}$$

The function g captures a very simple cognitive model in which people evaluate the generic by comparing observed prevalence to the known threshold. It is apparent that if the threshold θ were completely fixed, this model could account for neither context nor asymmetry effects. To set the stage for future Bayesian analyses, let us nonetheless explore the quantitative relation of a context-invariant threshold semantics to the data.

A context-invariant fixed-semantics

To begin our data analysis, we make no *a priori* assumptions about the (fixed, but unknown to us) value of θ, placing on it a uniform prior distribution: $\theta \sim U(0,1)$. We account for inattention and other irrelevant factors by including a probability $\phi_t \sim U(0,1)$ for each task that a given response is the result of uniform random guessing[4] (Lee & Wagenmakers, 2014). The inferred "guessing" parameter ϕ is the amount of data that would have to be attributed to random guessing in order for the fixed-threshold model of the generic to apply to the experimental data. In this sense, ϕ gives a coarse notion of model fit.

A joint analysis of the data from Exp. 1a and 1b, produces the expected extreme results. The inferred ϕ_{1a} is near 1, ϕ_{1b} has a more reasonable posterior mean of 0.3, and the inferred θ is near 100%. That is, the best interpretation of the data discounts the *truth conditions* data entirely and uses only the *implied prevalence* data to set the threshold, which results in interpreting the generic as a universal quantifier. A separate by-experiment analysis produces thresholds that are completely different for the two tasks. For the implied prevalence task, the threshold is again close to 100%. For the truth conditions task, the threshold is probably greater than 10% and less than 30%[5].

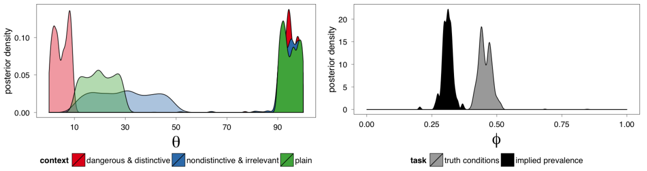


Figure 2: Analysis of the context-dependent fixed-semantics of a generic. Left: Inferred threshold for each experiment and context. Right: "guessing" parameter for each experiment.

A context-dependent fixed-semantics

In Exp. 1a, we replicated CBG's finding that the generic truth conditions vary by context. We reexamine our fixed-semantics model, now allowing for the possibility that θ could vary by context: $g(x,\theta_c)$. Again, we put uniform priors over each of the θ_c's. We are now in a position to examine how a truth-functional threshold of the generic would need to behave across these three contexts.

The results can be seen in Figure 2. The posterior distributions for ϕ_t's show that the amount of data that must be attributed to guessing to be consistent with this model of generics is still quite high: around 45% for the *truth conditions* task. Turning to θ_c, there is evidence for the variability of the generic threshold in the *truth conditions* task. In the *plain* context, the analysis suggests the threshold is somewhere between 0% and 30%, but it's unclear where exactly in that range the threshold should be. Critically, in the *dangerous and distinctive* context, the analysis infers a lower threshold, less than 10%. This matches with the earlier Null Hypothesis analysis. Finally, and most intriguingly, the analysis infers a third distinct threshold profile for the *nondistinctive and irrelevant* context. The inferred threshold is greater than 10%, but could be as high as 50%. This is an overall higher inferred threshold for the nondistinctive category; however, these results are inconclusive as to whether or not this threshold is actually different from the *plain* context.

The fixed-semantics model can also be evaluated by examining the posterior predictive distribution of responses. The posterior predictive distribution marginalizes over the inferred parameter values to produce predictions about what the data should look like given the cognitive model and the observed data. This is akin to fitting the parameters and is an important step in model validation as it shows what data is actually predicted by the model. Following directly from the inferred θ_c's, the model matches the ordering of the truth-conditions by context reasonably well. However, the model predicts an abrupt transition in endorsement rates for prevalences on either side of the threshold; the model is too dichotomous to match the human data. The correlation between the posterior predictive and the data is a moderate $r = 0.81$.

The fixed-threshold semantics model is not flexible enough to explain the observed data, but it has an even more serious

[4] Ideally, we would have ϕ be a function of participant (some participants guess more than others) and experimental condition (some conditions are more difficult or less constrained and invite more guessing). This is computationally too demanding when coupled with the more complex cognitive model explored later.

[5] This uncertainty results from the sparseness of our measure: participants were queried only at prevalence levels 10, 30, 50, 70, and 90%.

flaw: the variation of threshold by condition is postulated *a priori* in the data analysis, rather than accounted for by the cognitive model. That is, participants in our experiment must have some way of arriving at different thresholds for different tasks and conditions, which this model has no means to explain. For a more explanatory model we turn to the pragmatics of language understanding.

Reasoning about the threshold

The tasks in Exp. 1 are, fundamentally, language understanding tasks. We draw on recent work from probabilistic pragmatics to formalize our intuitions about how listeners arrive at interpretations of utterances. In particular, we draw on work from the Rational Speech Act (RSA) theory of language understanding. In this framework, a listener infers the meaning of an utterance by considering the thought-processes of a speaker whose goal is to be informative. Variants of this theory have provided computational explanations for a number of linguistic phenomena including scalar implicature, hyperbole, and argument evaluation (Kao, Wu, Bergen, & Goodman, 2014; Tessler & Goodman, 2014; Lassiter & Goodman, to appear).

We propose that the literal semantics of a generic sentence is in fact a threshold on prevalence, but listeners don't know the appropriate threshold and actively reason about it in context. A similar proposal has been made to explain gradable adjectives like *tall*. Lassiter and Goodman (2015) propose the meaning of an adjective like *tall* is a standard truth-functional meaning such that the object in question *is tall* if it has a height greater than the threshold θ_{tall}. The vagueness and context-sensitivity of scalar adjectives are accounted for by treating θ_{tall} as an unknown property of the language, and modeling the pragmatic listener as inferring this threshold.

The RSA model for generic interpretation, with the prevalence threshold as a variable "lifted" to pragmatic reasoning is specified by:

$$P_{L_0}(x \mid g, \theta) \propto g(x, \theta)P(x) \tag{2}$$

$$P_{S_1}(g \mid x, \theta) \propto P_{L_0}(x \mid g, \theta) \tag{3}$$

$$P_{L_1}(x, \theta \mid g) \propto P_{S_1}(g \mid x, \theta)P(x) \tag{4}$$

The literal content of the generic in Eq. (2) is identical to the fixed-threshold model in Eq. (1). However, it interacts with the prior distribution $P(x)$ over prevalence levels—the prior distribution of prevalence of a particular property across kinds.

Eq. (4) is a model of a listener (L_1) who has been told a generic statement. She assumes that, whatever the speaker (S_1) meant to communicate, the speaker was trying to be informative and that his goal was to communicate the prevalence x. She assumes the speaker in Eq. (3) knows θ and chooses an utterance to be informative to the literal listener (L_0). From this, the listener jointly infers both x and θ. We call this type of model a "lifted variable" model (lvRSA) because θ, traditionally thought to be part of the semantic content of the utterance (and thus perfectly transparent to all in the conversation), has been underspecified in the semantics but is locally fixed by pragmatic reasoning.

The prevalence prior $P(x)$ has a critical effect on the interpretation of the generic in this model. As a simplification, we posit a family of possible priors $x \sim \beta(\gamma, \delta)^6$. We hypothesize that the details of this prior (i.e. γ and δ) may differ according to the context in which a generic is used. For instance, when you know that a particular property is rare, a different distribution over categories is called to mind, than if the property is common. This results in different meanings for the generic. Below we infer appropriate prior parameters for each context from the behavioral data.

Following the advice of Degen and Goodman (2014), who investigated the relationship between dependent measures and speaker and listener roles in RSA, we will model the *implied prevalence* task as a pragmatic listener (L_1) task, but the *truth conditions* task as a pragmatic speaker task. We model the truth judgment with a speaker S_2 who is trying to convey the prevalence to a pragmatic listener, but can only produce the generic or its negation (i.e. yes or no to the truth of the generic):

$$P_{S_2}(g \mid x) \propto P_{L_1}(x \mid g). \tag{5}$$

The speaker in (5), like L_1, doesn't know the threshold, but knows that L_1 is thinking about it, and marginalizes over possible values: $P_{L_1}(x \mid g) = \sum_\theta P_{L_1}(x, \theta \mid g)$.

Results

To evaluate this model we perform a Bayesian data analysis similar to that used above. We infer the parameters of the prevalence prior, $\beta(\gamma, \delta)$, outside of the lvRSA cognitive model (but inside of the data analysis model), using uninformative priors: $\gamma_c \sim U(0,1)$, $\delta_c \sim U(0,5)$, $\phi_t \sim U(0,1)$, where $c \in \{DD, NI, P\}$ and $t \in \{$truth conditions, implied prevalence$\}$. We keep the data-analytic guessing parameters ϕ_t as a gross estimate of the proportion of responses not captured by our model of cognition.

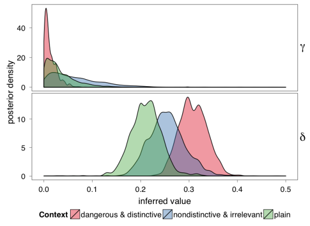


Figure 3: Posterior distributions of the hyperprior parameters used in lvRSA.

Inferred parameters The mean inferred values of ϕ_{1a} and ϕ_{1b} are about 0.08 and 0.05, respectively, a reasonable rate of "guessing" for participants on Amazon's Mechanical Turk, indicating that this cognitive model is doing a much better job of accounting for the signal in participants' responses.

Figure 3 shows the posterior distributions of the hyperprior parameters, γ and δ, for the lvRSA model. The posterior means for the γ's are well-ordered: $DD < P < NI$. This can

[6] For ease of interpretation, we are parametrizing the β distribution by its mean and concentration. To recover the canonical shape parametrization, use $\gamma\delta$ and $(1-\gamma)\delta$.

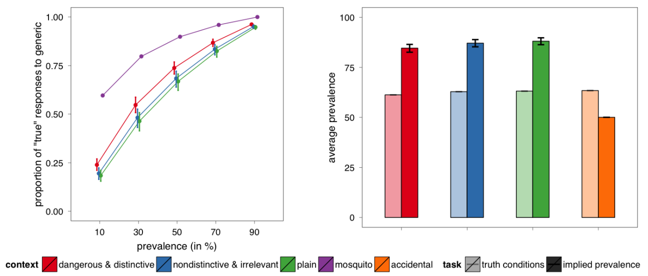


Figure 4: Posterior predictives of lvRSA for truth conditions (left) and asymmetry between dependent measures (right). Mosquitos and Accidental predictions use schematic priors (see description in Discussion).

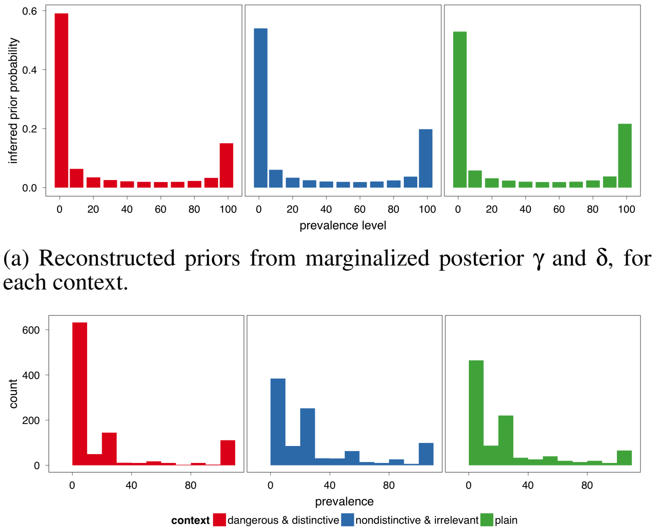
be directly interpreted as the mean prior prevalence for the three contexts: *dangerous and distinctive* properties are more rare than the other two types of properties. The δ's are much lower than 1 in each case, indicating bi-modal priors peaked at 0 and 1. The posterior means for the δ's are also ordered, suggesting that participants may treat variance of prevalence as higher in the *DD* context (or simply be more confused).

For further visualization, we marginalize over the posterior parameter values to reconstruct a canonical prior distribution over prevalence for each context. Figure 5a shows these prior distributions inferred from Exp. 1a & 1b data via the lvRSA model. Qualitatively, they are each bi-modal and the *DD* prior has a lower mean.

Posterior predictives The posterior predictions by lvRSA for the *truth conditions* task are shown in Figure 4. We can see that the model predicts graded endorsement rates for the generic as a function of prevalence—the model has some persistent uncertainty about the true value of the threshold. This same uncertainty is evident in the curves of Exp. 1a. With the inferred parameter values of the prevalence prior, the model also matches the differences in endorsement rates between context conditions. We reconstruct the curves of Figure 1 reasonably well; the model–data correlation is $r = 0.90$.

We use a similar data analysis strategy as we did for Exp. 1b to compare "average prevalence" between verification and interpretation. For the *truth conditions* task, we used the model's posterior probability of saying "true" at each prevalence level to simulate trials of the experiment as Bernoulli trials. We simulated 30 trials for each of 1000 imaginary subjects in this way. We then followed CBG's data analysis strategy (as recapitulated in Exp. 1b). The model gives a posterior distribution over prevalences, whose expectation we used to model the *implied prevalence* task. We find the model predicts the asymmetry between interpretation and verification of the generic (see Figure 4, right).

To see how this asymmetry is possible, consider again the inferred prevalence priors in Figure 5a. They are bimodal with peaks around 0% and 100%. This is consistent with the intuition that biological properties, such as the ones used by CBG, are properties either held by all of a category or none of a category. Since the semantics of the generic is underspecified (i.e. θ—the threshold of endorsement for truth judgement—is unknown), if θ falls anywhere in the range between 10%-90%, the most likely prevalence is going to be near 100%. Hence, in the *implied prevalence task*, the most likely inferred prevalence could be appreciably higher than one would expect from the *truth conditions* task.



(a) Reconstructed priors from marginalized posterior γ and δ, for each context.

(b) Priors elicited in Experiment 2.

Figure 5: Prior distributions over prevalence.

Often in Bayesian data analysis, the posterior distribution over parameters is hard to interpret in terms of observable phenomena. Our case is more transparent: if lvRSA were the correct model in this task, the prior distributions of prevalence for the three contexts should look like they do in Figure 5a. In particular, all three types of properties should have bimodal prior prevalence distributions, with a high probability that 0% of the kind have the property. Further, this left skew should be more pronounced for the *DD* properties relative to the *P* properties.

Experiment 2

Exp. 2 sought to test the prediction that the prior distribution of prevalence levels would be bimodal and vary by context.

Method

Participants We recruited 100 participants over Amazon's crowd-sourcing platform Mechanical Turk.

Procedure and materials Our procedure[7] was similar to Exp. 1b. On each trial, participants either read contextual information (*DD* or *NI*) or nothing (*plain*).

In addition to the contextual information, participants were presented with the following: "Listed below are 5 kinds of

[7] The experiment in full can be viewed at `http://stanford.edu/~mtessler/experiments/generics/cbg2010-replication/experiment/experiment-11.html`

animals, recently discovered." and asked the following question: "What percentage of each kind of animal do you think has [property]?" The experiment consisted of 6 trials, 2 from each context.

Results Experiment 2 recovered the shape of the inferred prior distributions predicted from the Bayesian analysis of the lvRSA model (compare Figure 5b to Figure 5a). Hartigans' Dip Test for Unimodality was highly significant for each of the prior distributions ($D = 0.054, 0.084, 0.0745$ for contexts *DD*, *NI*, *P*, respectively; p < 0.0001 for each), and thus the distributions are at least bimodal. The means of these three distributions are distinct and ordered as predicted (bootstrapped 95% confidence intervals in parentheses): $\mu_{DD} = 18.1\%(16.0, 20.2), \mu_P = 20.8\%(19.0, 22.5), \mu_{NI} = 25.7\%(23.7, 27.6)$. The medians of these three distributions were all significantly different from one another, evidenced by pair-wise Mann-Whitney U tests (*DD* vs. *P*: $W = 417452$; *DD* vs. *NI*: $W = 376180.5$; *NI* vs. *P*: $W = 548994.5$; all $p < 0.00001$). Finally, the distributions themselves were all significantly different from one another, by Kolmogorov-Smirnov tests (*DD* vs. *P*: $D = 0.185$; *DD* vs. *NI*: $D = 0.253$; *NI* vs. *P*: $D = 0.091$; all $p < 0.001$). In sum, the elicited prior distributions are all at least bimodal, have different central tendencies, and are all distinct.

Discussion

We have demonstrated the viability of a scalar semantics for generics when coupled with a sophisticated pragmatics. A lower-bound threshold on prevalence—the probability of the property given the category—is inferred as part of pragmatic interpretation, yielding vague and context sensitive meanings.

We formalized reasoning about the threshold in a lifted-variable Rational Speech Acts (lvRSA) model. This model predicted graded truth judgements and an asymmetry between truth and prevalence judgements. It also accommodated the role of context, explaining these effects as the result of variation in the prevalence prior. In Experiment 2, we verified that participants' beliefs about the prior on prevalence varied in this way. This provides evidence that the model we propose can account for many of the empirical phenomena associated with generics.

This model of generic interpretation makes further predictions for situations with qualitatively different prevalence priors. For example, if the prior distribution was unimodal, as in the case with incidental properties (e.g. broken legs), then the asymmetry between verification and interpretation could be dramatically reduced, cease altogether, or reverse (see Figure 4 for a prediction using the schematic prior in Figure 6). Indeed, CBG explored this possibility in one of their experiments using "accidental and disease states"; consistent with lvRSA, they found no asymmetry. Another example is a bimodal prior with a second peak at some low prevalence level (as opposed to a bimodal prior with a second peak at a high prevalence level, which we've focused on in this paper). This prior should describe rare properties that are not only rare *across kinds* but also rare *within kinds*. A canonical example of this is "West Nile Virus" in the generic "Mosquitos carry West Nile Virus". For a prior like this, the truth conditions for the generic would be relaxed at low prevalence levels, relative to the more common all-or-none priors. We see this behavior using a schematic prior (see Figure 4 for the truth conditions of the mosquito prior shown in Figure 6).

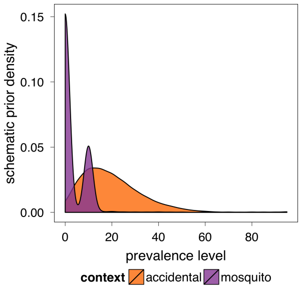


Figure 6: Schematic priors over prevalence of accidental properties and "mosquitos with West Nile Virus".

Generics are ubiquitous in natural language. It might seem paradoxical, then, that the semantics of generic statements are underspecified. Why should vague language get so much usage? One possibility is apparent in the lvRSA model: generic language provides interlocutors with the flexibility to convey rich meanings, which are easily understood in context. Generics are vague, but behaved and useful.

References

Carlson, G. N. (1977). *Reference to kinds in english* Unpublished doctoral dissertation. University of Massachusetts, Amherst.

Cimpian, A., Brandone, A. C., & Gelman, S. A. (2010). Generic statements require little evidence for acceptance but have powerful implications. *Cognitive science*, *34*(8), 1452–1482.

Cohen, A. (1999). Generics, Frequency Adverbs, and Probability. *Linguistics and Philosophy*, *22*.

Degen, J., & Goodman, N. D. (2014). Lost your marbles? the puzzle of dependent measures in experimental pragmatics. In *Proceedings of the thirty-sixth annual conference of the Cognitive Science Society.*

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

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition..

Kao, J. T., Wu, J., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*.

Lassiter, D., & Goodman, N. D. (2015). Adjectival vagueness in a bayesian model of interpretation. *Synthese*.

Lassiter, D., & Goodman, N. D. (to appear). How many kinds of reasoning? inference, probability, and natural language semantics. *Cognition*.

Lee, M. D., & Wagenmakers, E. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge: Cambridge University Press.

Leslie, S.-J. (2008, July). Generics: Cognition and acquisition. *Philosophical Review*, *117*(1).

Tessler, M. H., & Goodman, N. D. (2014). Some arguments are probably valid: Syllogistic reasoning as communication. In *Proceedings of the thirty-sixth annual conference of the Cognitive Science Society.*