Interpretation of Generic Language is Depends on Listener's Backgroud Knowledge

Anonymous CogSci submission

Abstract

Generic language, like "birds lay eggs" or "dogs bark" are simple and ubiquitou in naturally produced speech. However, the inherent vagueness of generics makes their interpretation highly context dependent. Building on work by Tessler & Goodman (in press) showing that generics can be thought of as inherently relative (i.e. more birds lay eggs than you would expect), we explore the consequences of different implied comparison categories on the interpretaiton of novel generics. In Experiment 1, we manipulated the set of categories salient to a listener by directly providing them the comparison sets. In Experiment 2, we collected participants' demographic information and used these naturally occurring differences as a basis for differences in the participants' comparison sets. Results from both studies confirmed our hypothesis that the prevalence of a feature in different comparison categories changes people' estimatite of the feature prevalence in novel categories. ONE MORE SENTENCE HERE ABOUT IMPLICATIONS

Keywords: generics; semantics; meaning; learning; Bayesian inference

Introduction

Generic language like "birds lay eggs" is a simple, highly frequent way of transmitting information in everyday speech (Gelman, Goetz, Sarnecka, & Flukes, 2008; Gelman & Tardif, 1998). Generics are distinct from statements about particular referents "e.g. that bird lays eggs"; they transmit information about *categories*. Indeed, a large body of research has documented the power of generic language in adults' and children's inference about familiar and novel categories (e.g., Cimpian, Brandone, & Gelman, 2010; Cimpian & Markman, 2011; Rhodes, Leslie, & Tworek, 2012). Despite their ubiquity, generic statements defy a straightforward definition in threshold semantics (i.e. they do not specify a fixed prevalence rate). While people generally agree that "birds lay eggs," this does not mean that "all birds lay eggs (100%)" nor does it mean that "most birds lay eggs ($> 50\%$)"–male birds, and young female birds do not. Similarly, "birds lay eggs" cannot mean "some birds lay eggs ($> 0\%$)," because it is true while "birds are female" is not.

Recent work from Tessler & Goodman (in press) shows that generics can be understood through the lens of Gricean pragmatic inference (Grice, 1975). Their key insight is that generics can be interpreted as statements about relative prevalence. If a speaker makes a vague statement like "birds lay eggs," but listeners assume that they are cooperatively intending to be informative, they can infer that the speaker means something like "birds are *more likely than you would have expected* to lay eggs." This formulation leaves open two questions: (1) how much more likely does a speaker mean, and (2) what did the listener expect? Tessler and Goodman answer the first question by showing that listeners do not need to resolve this ambiguity directly, but can instead integrate over all prevalence rates that would make the speaker's statement true. In a series of experiments with both familiar and novel generics, Tessler & Goodman (in press) show that people's judgements about prevalence rates following a generic statement are described by a rational model pragmatic inference (Frank & Goodman, 2012).

We take up the second question: How do listeners arrive at their prior expectations? One possibility is that implicit in a generic statement is a set of reference categories, i.e. "birds lay eggs" means "relative to relevant comparison categories, birds are more likely to lay eggs." The listener's interpretation of a generic, then, should depend on the set of categories they consider relevant. That is, "feps are friendly (relative to puppies)" should lead to a much different estimate of the prevalence of friendliness in feps than "feps are friendly (relative to squirrels)."

We test this prediction in a two experiments in which people learn about novel categories through generic language. In the first, we manipulate the implied comparison category directly and show that people's judgments about the prevalence of a feature in a novel category tracks the prevalence level of the implied category. In a second experiment, we show the influence of implicit comparison categories without manipulating them. Here we leverage prior work showing that people's estimates about the prevalence of preferences and beliefs in others are egocentrically biased towards the prevalence of those preferences and beliefs in their local communities (Ross, Greene, & House, 1977). Together, these studies highlight the fundamentally relative way in which even simple generic statements are interpreted, and point towards a potential source of misunderstanding and errors in learning that can arise from well-intentioned communication.

Experiment 1

Two conditions were tested in this experiment - (1a) a baseline survey, where we asked participants to estimate prevalence rate for familiar category-feature pairs, and (1b) a novel category survey, where participants were introduced to a

novel category along with a familiar comparison category, then they were shown a generic containing a novel category and a familiar feature and asked to estimate the prevalence rate of the feature within the novel category.

Method

Participants A total of 300 participants were recruited on Amazon Mechnical Turk, 150 participated in Experiment 1a, and 150 participated in Experiment 1b. Each participant gave informed consent at the start of the Experiment and was paid **AMOUNT** in compensation. Participants were excluded from the final sample if they did not pass an attention check at the end of experiment (5 in Experiment 1a, and 36 in Experiment 1b) yielding a final sample of 145 participants in 1a and 114 participants in 1b.

Design and Procedure Three features (friendly, tasty, and heavy) were chosen, and, for each feature, we chose three categories that were relevant to the feature that will elicit different levels(low, medium, high) of prevalence estimation. Every participant answered questions about each of three features and, for each feature, one randonly-selected category. The order in which the features appeared in the survey was randomized, and each participant was tested on only one category from each of the predetermined prevalence levels.

In Experiment 1a, participants were first shown a generic for each of the category and feature pair. Then they were asked to first evaluate the truth condition of the generic by answering a forced-choice True or False question, and to estimate the proportion of the feature within the given category. Participants recorded their responses to the estimation question on a scale slider, ranging from 0% to 100%.

In Experiment 1b, participants answered similar questions but about novel categories. Participants were first told that they are visiting three new countries, and people from there will introduce them to things in their respective countries. Each novel category was first introduced by reference to one of the baseline categories from Experiment 1a (e.g. "Feps are like puppies"). Then, they were given a novel generic using the same features about this novel category (e.g., "Feps are friendly"). All other aspects of the design were identical.

After completing these questions, participants were given an attention check to ensure that they had read and engaged with the stimuli. In Experiment 1a., we asked participants to choose the three features we asked about in the survey. In 1b., we asked participants to select the three novel categories that were mentioned in the survey.

Results and Analysis

Figure 1 shows participants' mean prevalence judgments across conditions in both conditions. In both Experiments, participants' judgments about the prevalence of the target feature in each category increased as predicted from low to medium to high. In addition, judgments made about the novel category were on average higher than the judgements for the corresponding category made by participants in the baseline

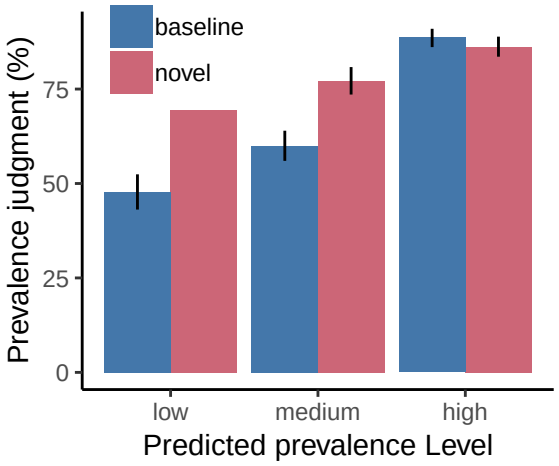

Figure 1: Prevalence judgments of participants in the Baseline (1a) and Novel (1b) conditions. Error bars indicate 95% confidence intervals computed by nonparametric bootstrapping

condition, although this difference was not apparent in the high prevalence condition. We confirmed these predictions statistically using a mixed-effects logistic regression, predicting participants' judgments from condition and prevalence level and their interaction, with random effects of subject and feature (`prop ~ condition * level + (1|subj) + (level|feature)`). This model revealed a significant effect of level, with both medium ($\beta = 0.12$, $t = 5.41$, $p < .001$) and high ($\beta = 0.41$, $t = 17.97$, $p < .001$) levels of apriori prevalence receiving higher prevalence judgments, a significant effect of condition ($\beta = 0.26$, $t = 9.76$, $p < .001$), and a significant interaction between the two for both medium (($\beta =$ -0.09, $t =$ -2.63, $p = .009$) and high levels ($\beta =$ -0.29, $t =$ -8.27, $p < .001$) indicating that the change in prevalence levels was largest for the lowest apriori level.

Discussion

NOT THE REAL TEXT WE WANT We found the effect. Maybe some ceiling effects on prevalence? This is great because it shows we can manipulate the effects, but it's pretty explicit. Next we show we can do it with apriori estimates. First we get baseline estimates for demographic groups, then we show that these predict judgments about novel generics.

Experiment 2

Method

Experiment 2 is composed of two experiments, where experiment 2a includes a baseline survey of 15 questions about people's habits, and a simulation to select a smaller set of 6 questions to in experiment 2b. Participants were asked to make estimations on the prevalence rate of each habit among people. Experiment 2b used the results from the previous simulation to finalize the 6 questions to be asked. In experiment 2b, participants were introducted to six novel countries

and people in those countries, and were asked to make prevalence estimations for each habit among people in the foreign countries. In both experiments, we collected data on particpants demographic information, including their gender, age, political ideology score, and zip code. We later used the demographic information to analyze the naturally occuring differences in particpants' comparison sets when making prevalence estimations.

Participants For Experiment 2a. baseline survey, we recruited 968 adult participants from Amazon Mechanical Turk. 726 participants (Mage = 37 years, SD = 11.92) passed the final attention check and their responses were included in the following analysis. For Experiment 2b novel category survey, 400 adult participants were recruited from Amazon Mechanical Turk. 317 participants (Mage = 35 years, SD = 11.47) passed the final attention check and their responses were included in the analysis. Participants from both conditions were asked a series of demographic questions, including gender, age, political ideology score, and zip code. Participants recruited were U.S. citizens over 18 years old and received monetary compensation for their work.

Exp 2a. Baseline Survey Participants in this condition were shown a series of 15 questions. Each question asked them to make an estimate the percentage of people having a certain habit (see Table. 3). Participants responded by choosing a number from 0 - 100% using a slider bar. The order in which the questions appeared was randomized. Six questions were chosen from the previous fifteen questions in Experiment 2a to be asked in Experiment 2b. The selection was based on running simulation (n = 100) on baseline survey responses from participants to all 15 questions and we selected a combination of 6 questions that has large t values.

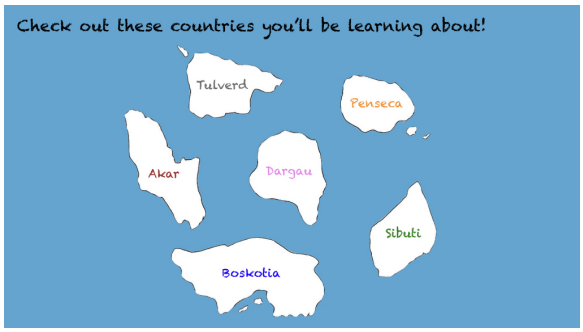

Figure 2: Novel category survey map prompt

Exp 2b. Novel Category Survey Participants in this condition were first shown a map of 6 imaginary countries and their corresponding novel names (see Fig. 2) along with a prompt. They were told that people from these six countries will introduce some habits of people from their countries to them. They were then introduced to generic statements about each of the country and a habit selected from the simulation. After reading the generic statement, participants were then asked to make an estimate the percentage of people having a certain habit (see Table. 3). Participants responded by choosing a number from 0 - 100% using a slider bar. The order in which the questions appeared was randomized. The code for analyze this experiment is preregistered.

Results

We calculated the pearson correlation of average mean responsess across demographic groups between experiment 2a. baseline and 2b. novel for all 6 questions selected via simulation. The average mean responses between the two experiments are positively correlated for each question, with p-value of the t-test equals to 0.002. The highest correlation among the six questions was 0.688 for the feature "like to cook at home", and the lowest correlation was 0.216 for "go to the gym". We further examinzed the within-demographic group differences in mean responses between the baseline and novel experiments by questions (see Fig.2). We divided the demographic groups into 8 subgroups (2 age, 2 gender, 2 political leaning). The bins for dividing age and political ideology scores in experiment (2a) were based on the mean age and mean political ideology scores in experiment (2b). Each dot in the sub-plot in Fig.2 is a unique demographic group. Across all six questions, the mean responses from each of the 8 demographic groups for the novel conditions (shown in y-axis) were consistently higher than the mean responses for the baseline condition. The results from experiment 2 showed that participants from different demographic groups make different prevalence estimations within each question, and their responses were on average higher for novel conditions than for baseline conditions. Participants across demographic groups also respond to different questions differently.

Acknowledgements

Place acknowledgments (including funding information) in a section at the end of the paper.

References

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.

Cimpian, A., & Markman, E. M. (2011). The generic/nongeneric distinction influences how children interpret new information about social others. *Child Development*, *82*(2), 471–492.

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

Gelman, S. A., Goetz, P. J., Sarnecka, B. W., & Flukes, J. (2008). Generic language in parent-child conversations. *Language Learning and Development*, *4*(1), 1–31.

Gelman, S. A., & Tardif, T. (1998). A cross-linguistic comparison of generic noun phrases in english and mandarin. *Cognition*, *66*(3), 215–248.

Grice, H. P. (1975). Logic and conversation. *1975*, 41–58.

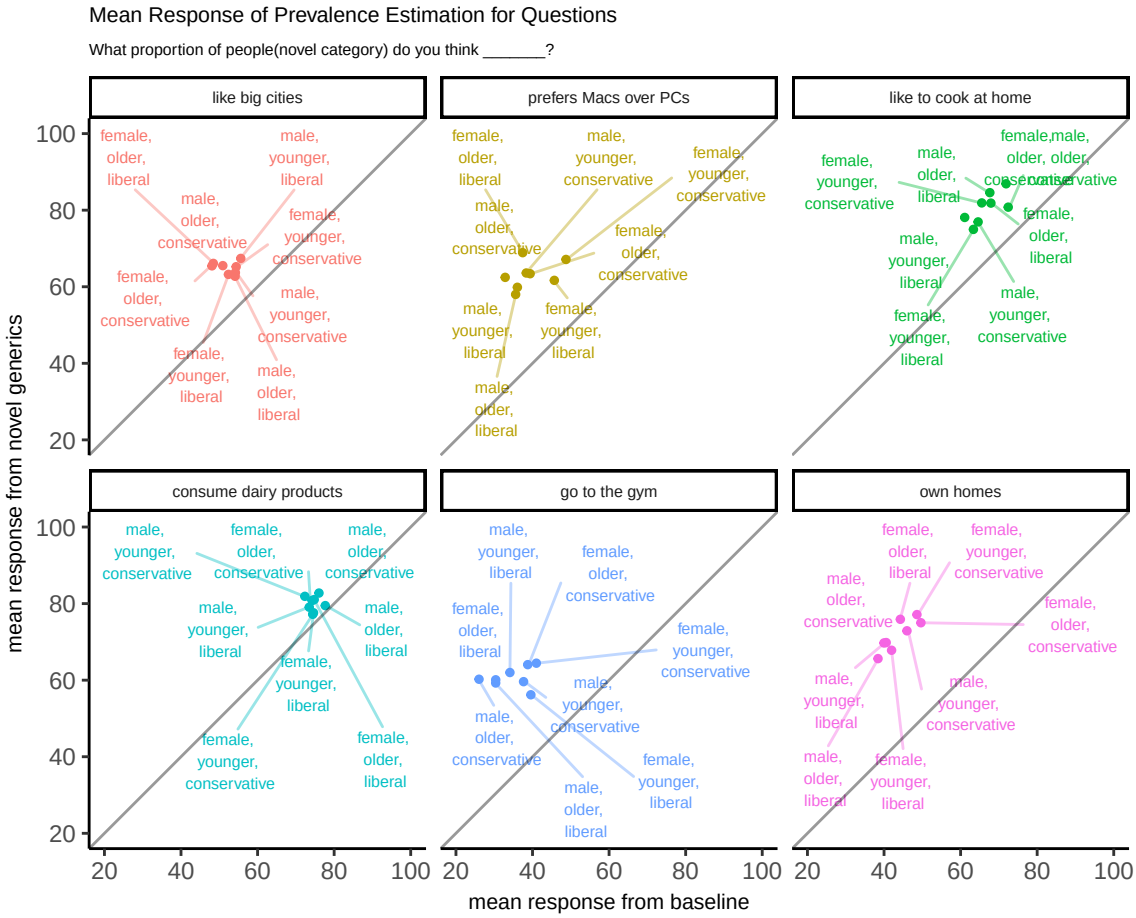

Figure 3: Novel vs. baseline category mean response

Rhodes, M., Leslie, S.-J., & Tworek, C. M. (2012). Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences*, *109*(34), 13526–13531.

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, *13*(3), 279–301.

Tessler, M. H., & Goodman, N. (in press). The language of generalization. *Psychological Review*.

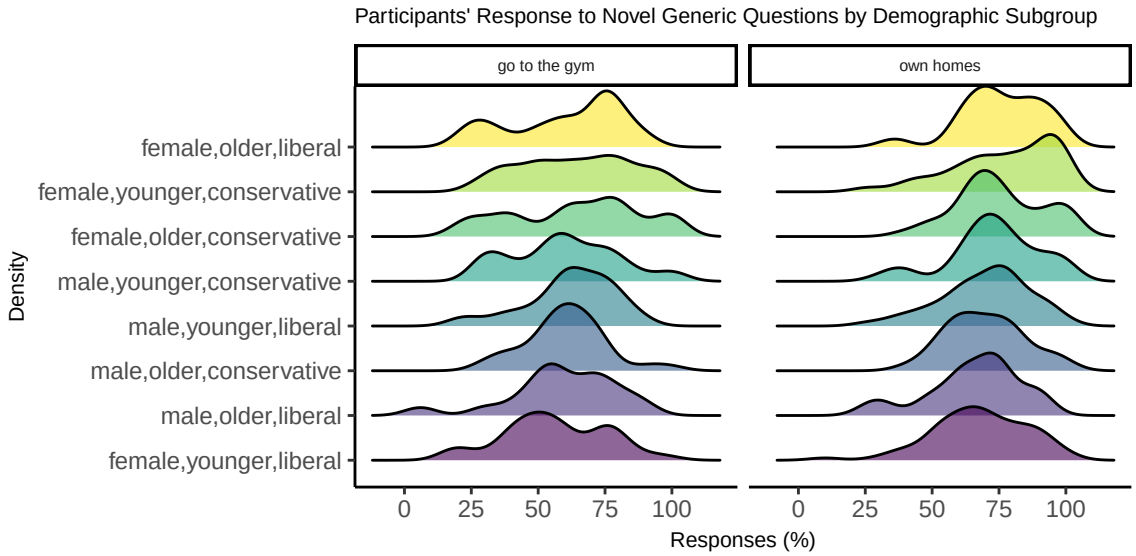

Figure 4: Novel Condition Prevalence Estimation by Demographic Groups