Meanings of Generic Language are Sensitive to Listeners' Backgroud Knowledge

Anonymous CogSci submission

Abstract

Children reguarly hear and produce generic languages during their daily interactions with their parents. Generics such as "birds lay eggs" or "dogs bark" provide kind-based information about a category (e.g., birds) and its characteristic features (e.g., lay eggs, bark.) This function allows speakers to successfully communicate generalizable knowledge about the distribution of certain features within a given category without the constrain of referring to objects only in the here-and-now. However, the meaning of generics can be ambiguious and highly context-dependent. Without any quantificational information, the interpretation of a generic depends largely on a listener's own world knowledge. In a series of 2 experiments on Amazon Mechanical Turk, we investigated how differences in listeners' comparison sets may lead them to make systematically biased inferences of prevalence for a given feature in novel categories. 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.

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

Introduction

Generic languages, such as "birds lay eggs," are commonly used in parent-child conversations (Gelman & Tardif, 1998.) Different from sentences such as "that bird lays eggs" and "some birds lay eggs," previous research has shown that generics were more likely to elicit people's responses about kinds rather than specific instances of a kind (cite Gelman; Rhodes; Cimpian; Tessler). Conveying general information about a given category and its property, generics are part of our vernacular since a young age. Parents from English-speaking and Mandarin-speaking families regularly produce generics when talking with their children (Gelman & Tardif, 1998). Adults also utilize generics in their daily speech and writing (Master, 1987). However, despite its appearance in both speech and text, which may suggest it providing a privileged path to information transmission (easily understandable, highly accurate, etc.), the meaning of generic language is ambiguous.

The composition of a generic, in comparison to other forms of statements, contributes to its vagueness in meaning. For instance, consider the bare plural noun phrase "birds" in "birds lay eggs", it makes a claim about the bird category that underspecifies the prevalence rate of the feature *lay eggs* in birds. "Birds lay eggs" is neither a statement about all members of that category laying eggs, nor half of the members, nor any specific proportion. If one were to evaluate the meaning (or truth value) of a generic and think carefully the criterion for which "birds lay eggs" would be true, one would realize that, for birds to lay eggs, they have to be adult, female, and healthy. These criterion immediately restrict the proportion of birds that lay eggs to be about 50% if not less. Thus, a reasonable interpretation of this generic relies heavily on a listener's prior knowledge of the category "birds" and of the activity "lay eggs". Upon hearing "lay eggs", a listener may infer that female birds are the subject under discussion even though "female" is not explicitly mentioned so as to not mistakenly infer that all members of the category birds lay eggs. In contrast, quantifiers such as "some", "most", and "all" limits the interpretation of a sentence to a certain range of responses by providing a filtering criteria for prevalence distribution in members of a category. In the absence of quantifiers, the meaning of generics become more context-dependent and sensitive to listeners' prior knowledge. People can generate quite different readings of the same generic, depending on their understanding of the underlying distribution of the category and property in a given generic.

unfinished Recent work has proposed to formalize the meaning of generic expressions as conditional probability (Tenenbaum, Kemp, Griffiths & Goodman, 2011; Frank & Goodman, 2012; Tessler & Goodman, 2017). Instead of specifying a fixed prevalence threshold for each category-feature pair, an interpretation for a generic can be seen as derived from listeners utilizing their probabilistic world knowledge upon hearing the statement.

Experiment 1

Method

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.

Participants For condition 1a. baseline survey, we recruited 150 participants from Amazon Mechanical Turk. 145 participants passed the attention check question and their responses were included in the following analysis. For condition 1b. novel category survey, 150 participants were recruited from Amazon Mechanical Turk. 114 participants passed the attention check question and their responses were included in the analysis. Participants recruited were U.S. citizens over 18 years old and received monetary compensation for their work. We inserted a Unique Turker Id scirpt in our survey designs to ensure that each worker on Turk is allowed to take only one survey.

Condition 1a. Baseline Survey In condition 1a., 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. These categories were later provided to the participants in condition (1b) as comparison sets for their inferences of the novel category. 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 the survey, 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%.

Condition 1b. Novel Category Survey In condition 1b., the same three features (friendly, tasty, and heavy) were used. Different from condition (1a), for each feature, we introduced participants to a novel category using made-up words. Parctipants were also provided with categories from condition 1a. as a reference set in the form of "Feps(the made-up word) are like puppies(a baseline category)". The paired baseline category was randomly selected from the three possible categories for each feature. Same as 1a., every participant answered questions about all three features, and one randomly selected comparison category for each feature. The order in which the features came in the survey was again randomized, and each participant was tested on only one category from each predetermined response range (1 low, 1 medium, and 1 high). In the survey, 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 embedded in a generic (e.g., "Feps are friendly") when first introduced, followed by an explicit of comparison set (e.g., "Feps are like puppies"). After which, participants were asked to estimate the proportion of the feature within the novel category. Participants recorded their responses to the estimation question on a scale slider, ranging from 0% to 100%.

Attention check question For both conditions, we designed an attention check question in the end of the survey. For 1a., we asked participants to choose the three features we asked about in the survey. For 1b., we asked participants to select the three novel categories that were mentioned in the survey.

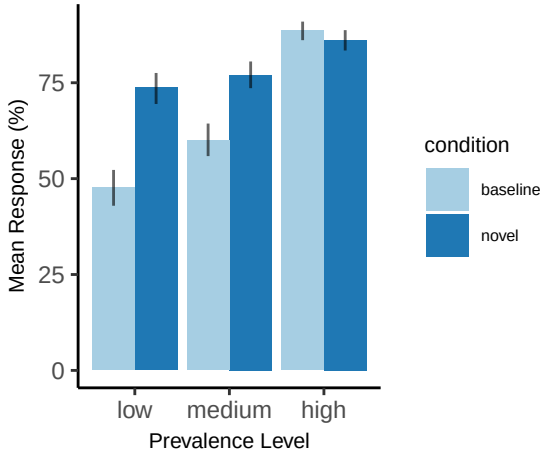

Figure 1: Baseline vs. Novel Condition Prevalence Estimation

Results

Participant responses in both conditions increased by prevalence level (Fig. 1). In the baseline condition, participants' responses by prevalence levels were consistent with our prior grouping of the categories across features. For the low, medium, high prevalence level, the respective mean response are: M = 47.7 (CI = 43.1-52.4), M = 60.0 (CI = 55.9-64.1), and M = 88.7 (CI = 86.1-91). In the novel category condition, the respective mean response are: M = 73.8 (CI = 69.8-77.5), M = 77.1 (CI = 73.4-80.6), and M = 86.1 (CI = 83.4-88.7). Both the baseline and novel conditions exhibited a trend of increase in prevalence estimation by prevalence levels. Furthermore, consistent with our model's prediction, the mean responses of novel conditions in both low and medium prevalence levels were higher than the baseline condition. However, the novel condition's mean response for the high prevalence level is lower than the baseline condition's by 2.6, and their confidence intervals overlapped.

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.

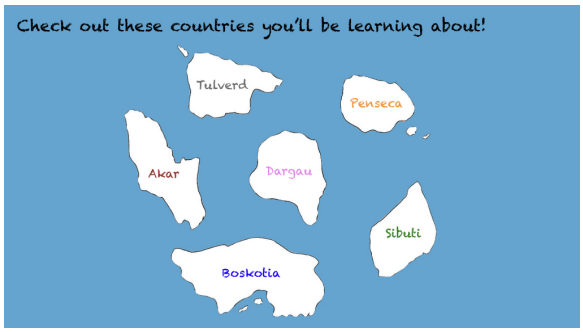

Figure 2: Novel category survey map prompt

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

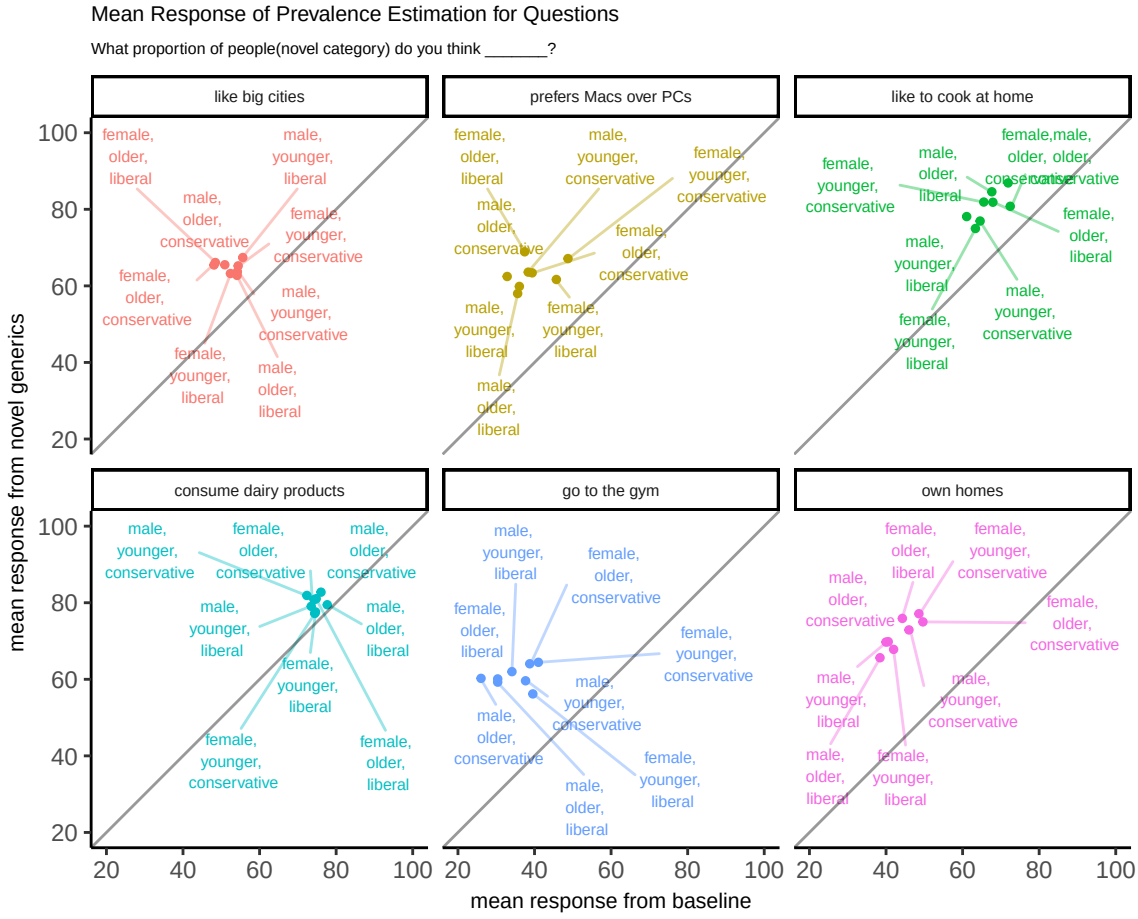

Figure 3: Novel vs. baseline category mean response

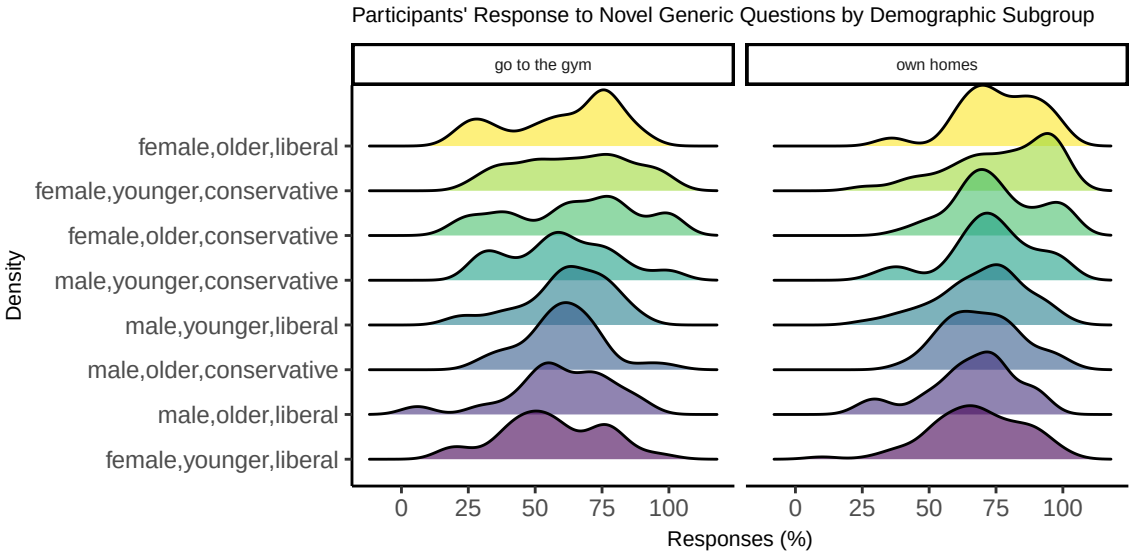

Figure 4: Novel Condition Prevalence Estimation by Demographic Groups