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

Anonymous CogSci submission

Abstract

Generic language, like "birds lay eggs" or "dogs bark" are simple and ubiquitous 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' estimations of the feature prevalence in novel categories. These results, highlighting how context-sensitive interpretations of generic language are to listeners' prior knwoledge, suggest a possible source for *well-intentioned* miscommunications, where conversational partners are cooperative during a discourse but are led by their different backgrounds to make dissimilar inferences of the same statement.

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 of 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 series of four experiments in which people learn about novel categories through generic language. In the first two experiments, 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 Experiments 3 and 4, 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

In this experiment, we asked participants to make prevalence judgments for familiar category-feature pairs. Partcipants' prevalence estimations will be used as a measure of baseline prevalence ratings for each feature in a given category. These familiar categories will be provided to a different group of participants in Experiment 2 as comparison categories.

Method

Participants. 150 participants were recruited on Amazon Mechnical Turk. Each participant gave informed consent at the start of the Experiment and was paid 10 cents in compensation. Participants were excluded from the final sample if they did not pass an attention check at the end of experiment (5), yielding a final sample of 145 participants.

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.

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 then to estimate the proportion of the feature within the given category. Participants gave their responses to the estimation question on a scale slider, ranging from 0% to 100%.

After completing these questions, participants were given an attention check to ensure that they had read and engaged with the stimuli. Here, we asked participants to choose the three features we asked about in the survey from a larger set of features.

Results and Analysis

Participants' mean prevalence judgments about the target feature in each category increased as predicted from low to medium to high (as shown by the baseline condition in Figure 1). We confirmed this prediction statistically using a mixed-effects logistic regression, predicting the participants' judgments from prevalence level, with random effects of subject and feature (`prop ~ level + (1|subj) + (level|feature)`). This model revealed a significant effect of level, with both medium ($\beta = 0.12$, $t = 5.02$, $p < .001$) and high ($\beta = 0.41$, $t = 16.59$, $p < .001$) levels of *a priori* prevalence receiving higher prevalence judgments.

Furthermore, the number of participants who evaluated the generic as true also varied across conditions ($n_{low} = 78$, $n_{medium} = 117$, $n_{high} = 144$), with an increased number of participants evaluating true for generics that contained categories from higher prevalence levels (medium and high). Participants who evaluated the generic as true for categories from the low prevalence level also on average made higher estimation of prevalence ($\mu_{low} = 67.75$, $ci_{lower} = 63.22$, $ci_{upper} = 72.04$) comparing to all participants for categories in the same prevalence level($\mu_{low} = 47.72$, $ci_{lower} = 42.96$, $ci_{upper} = 52.34$). We then used a mixed-effect logistic regression to predict participants' judgments from prevalence level and their true or false evaluation of the generic as well as their interaction, with random effects of subject and feature (`prop ~ level * response + (1|subj) + (level|feature)`). This model revealed a significant effect of the true or false evaluation, with the true evaluation($\beta = 0.4$, $t = 12.83$, $p < .001$) leading to higher prevalence judgements, a significant effect of the aprior high prevalence level ($\beta = 0.8, t = 4.39, p < .001$) receiving higher prevalence judgments, and a significant interaction between the true evaluation and high prevalence level ($\beta = -0.58$, $t = -3.13$, p .002).

Discussion

The results from Experiment 1 confirmed our hypothesis that participants' estimations of feature prevalence for a category in a given generic varied by the predetermined prevalence level of comparison categories. Participants who evaluated true for the generics gave similar estimations on avaerage for categories from low and medium prevalence levels while providing a significantly higher mean prevalence rating for the categories in the high prevalence level. One possibility is that participants who agreed with the generic are more likely to give a relatively higher prevalence ratings, resulting in similar mean prevalence estimations for categories from the low and medium prevalence levels. However, participants' prevalence judgmenet were still highly sensitive to the comparison category (from aprior low and medium prevalence levels or the high prevalence level) rather than a fixed threshold. Furthermore, the chosen category-feature pairs successfully elicited participants' responses across diffferent prevalence levels. Next, in Experiment 2, we introduced participants to novel categories along with the same set of familiar comparison categories and features. We predict that participants' estimation of a given feature in a novel category will be sensitive to its respective familiar comparison category and the aprior prevalence levels.

Experiment 2

Experiment 2 included 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. 150 participants were recruited through Amazon Mechanical Turk. Each participant gave informed consent at the start of the Experiment and was paid 10 cents in compensation. Participants were excluded from the final

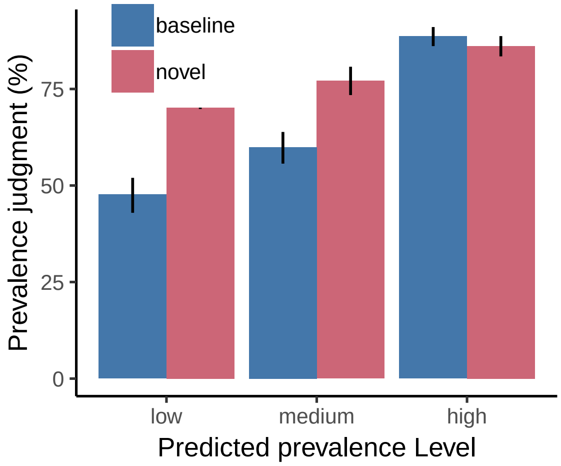

Figure 1: Prevalence judgments of participants in the Baseline (1a) and Novel (1b) conditions. Error bars indicate 95% confidence intervals computed by nonparametric bootstrapping

sample if they did not pass an attention check at the end of experiment (36), yielding a final sample of 114 participants.

Design and Procedure. The same sets of features and categories from Experiment 1 were included in Experiment 2. Additional, three novel categories were introduced (one novel category per feature). Participants answered similar questions as the baseline survey in Experiment 1 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 1 (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. 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 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

The results from Experiments 1 and 2 confirmed our hypothesis that particpants’s judgements about the prevalence of a feature in a novel category tracks the prevalence level of the implied category. The difference in estimation on average between baseline and novel conditions was significant for both the aprior low and medium levels, with the novel conditions’ estimation higher than the baseline condition, while the differences between the two conditions for the high prevalence level was not apparent. One possibile explanation for the small difference may be that there exists a ceiling effect for the estimations of high prevalence categories, considering that the upper limit for any prevalence rating was 100%.

Next, in Experiments 3 and 4, we further explored the effect of differences in comparison set without provding participants explicit comparison categories. We instead used naturally occuring differences among participants by collecting participants’ demographic information. First, we obtained baseline prevalence estimations of a group of features for different demographic groups in Experiment 3, and showed that the responses in Experiment 3 predicted the prevalence judgements about novel generics in Experiment 4.

Experiment 3

In Experiment 3, we collected prevalence judgments for a set of 15 habits, activities, and preferences along with demographic information. Our goal was to use this set of questions to find features on which prevalence rate estimates varied across demographic groups.

Method

Participants. 968 participants were recruited on Amazon Mechanical Turk. Each participant gave informed consent at the start of the experiment and was paid 50 cents in compensation. Participants were excluded from the final sample if they did not pass an attention check at the end of experiment. 195 were excluded for failing the attention check. We further excluded participants who self-identified as gender nonconforming (3) or participants over 60-year-old (44) as the data for these two categories were too sparse for our purpose of analysis, yielding a final sample of 726 participants ($M_{age} = 36$, $sd = 10.05$).

Design and Procedure. Fiftheen features were chosen to be included in the survey. Each feature was either a familiar habit or activity people might participate in their daily lives (e.g., like to cook at home, go to the gym, consume dairy products). Every participant was asked to estimate what proportion of people who engaged in that particular habit or activity. The order in which the features appeared in the survey was randomized. After making these judgments, participants were asked to report their age, gender, political ideology score, and zip code. Participants recorded their responses for the estimation question on a scale slider, ranging from 0% to 100%. For the collection of demographic information, participants typed their age in a text box, selected one choice from either "female", "male", or "other/non-conforming" for the gender question, recorded their political ideology score on a slider ranging from 1 (most liberal) to 7 (most conservative), and typed their zip code in a text box. After completing these quetsions, participants were given an attention check to ensure that they had read and engaged with the stimuli. We asked participants to select four features we asked about in the survey.

Results and Analysis

Although we collected a large number of judgments (10,890), we struggled to detect reliable demographic difference in judgments due to non-independence of group membership that made estimation of individual group effects difficult (e.g men were more likely to be conservative). However, we were underpowered to treating each conjunction of groups (e.g. male, conservative, older) as an independent demographic level. Further, peopel in different demographic groups had different biases in the use of the 0% to 100% scale (e.g. men gave higher values for every question). Nevertheless, we wanted to find a subset of the most demographically variable questions for use in Experiment 4. We decided that from the total set of 15, we would choose the 6 for which there was the most reliable variability across demographic groups. In order to find the best subset, we turned to simulation.

For each step of the simulation, we would first select 6 of the 15 questions. We divided our data randomly into two halves, and computed the mean judgement for each question given by each conjoined demographic group in each half (e.g. men gave higher values for every target question). We then scaled the responses in each group across question to remove biases in scale use, and finally computed the correlations in demographic groups' judgments for each individual question across the two halves of the data. We computed this correlation for 1000 random samples for each possible combination of 6 of 15 questions. We then chose the set of 6 questions for which the correlations across two random splits of the corpus was highest. Figure 2 shows two questions that were characterized by this stable regularity. These features show both the biases in scale use that these demographic groups brough to the task (e.g. female older liberals gave high values in general), but also the relative variability across questions (relative to their judgments about what proportion of people go to the gym, female young conservatives give lower values than we would expect given the values they typically give).

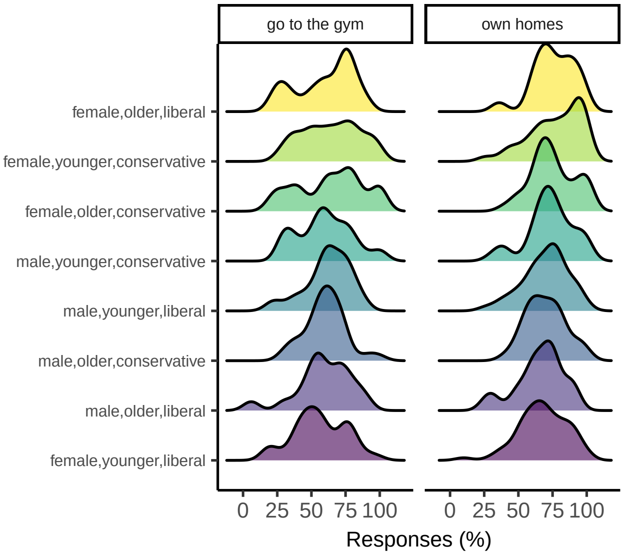

Figure 2: Novel Condition Prevalence Estimation by Demographic Groups

Because we developed this simulation method after having observed the data, we pre-registered this set of questions and a full analysis plan on the Open Science Framework. We predicted that a new set of participants, dvidied into the same demographic groups, would give judgments about the prevalence of these 6 habits and activities in a novel group of people that was correlated with their judgments about people they knew.

Experiment 4

Method

Experiment 4 used the simulation result in Experiment 3 to finalize a set of 6 questions to be asked in a novel category survey. In the novel category survey, participants were introducted to six novel categories, then they were shown a generic containing a novel category and a familiar feature, and asked to make prevalence estimations for each habit among people in the foreign countries. Participants were asked to provide some demographic information, including their gender, age, political ideology score, and zip code.

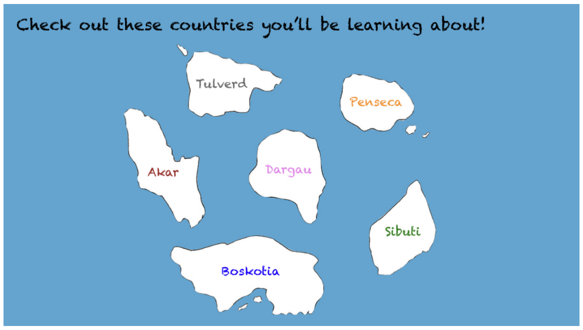

Figure 3: Novel category survey map prompt

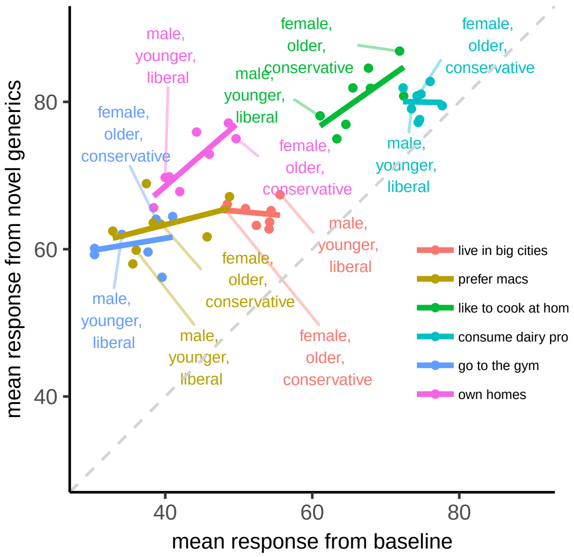

Figure 4: Novel vs. baseline category mean response

Participants. 398 participants were recruited on Amazon Mechanical Turk. Each participant gave informed consent at the start of the exerpiment and was paid 0.2 dollar in compensation. Participants were excluded from the final sample if they did not pass an attention check at the end of experiment. 63 were excluded for failing the attention check. We further excluded participants who self-identified as gender non-conforming (2) or participants over 60-year-old (16), yielding a final sample of 317 participants (M_{age} = 33, sd = 9.48).

Design and Procedure. Participants in this condition were first shown a map of 6 imaginary countries and their corresponding novel names (see Figure 3) 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 or activity. Participants responded by choosing a number from 0% - 100% using a slider bar. The order in which the questions appeared was randomized. After completing these questions, participants were given an attention check to ensure that they had read and engaged with the stimuli. We asked participants to select four features that were mentioned in the survey. The code for analysis of this experiment is preregistered.

Results

We calculated the pearson correlation of average mean responsess across demographic groups between Experiments 3 and 4 for each of the 6 questions selected via simulation. The average mean responses between the two experiments are signficantly correlated for each question ($\beta = 0.46$, $t = 5.58$, $p = .003$). The highest correlation among the six questions was 0.69 for the feature “like to cook at home”, and the lowest correlation was 0.22 for “go to the gym”. We further examined the within-demographic group differences in mean responses between the baseline and novel experiments by questions (see Fig 4). The demographic groups were divided into 8 subgroups (2 age bins × 2 gender bins × 2 political ideology bins). The bins for dividing age and political ideology scores in Experiment 4 were based on the mean age and mean political ideology scores in Experiment 3. Each dot in a group of same coloured dots in Figure 4 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 Experiments 3 and 4 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.

One of our motivations for using demographic groups as a proxy for different clusters of comparison sets is that people may sample their comparison categories from localized communities, where members of the same demographic groups are more likely to be in vincity of each other. However, the results from these different demographic groups may also be highly confounded, since it is more likely that people do have interactions with others who are not exactly in their specified demographic groups (e.g., a female older liberal person interacts with a male younger liberal person in her daily life). Moreover, the results from Experiments 3 and 4 may be confounded with people’s ingroup/outgroup biases due to the selection of features, which may trigger people to use their ingroup versus outgroup distinction to make estimations depend on whether they were making judgments on the baseline *people* or the novel catgeory (people from other novel countries). Our results showed that people systematically overestimate activities of others from outgroups upon hearing a novel generic but their estimations make reference to their ingroup beliefs.

General Discussion

In Experiments 1 and 2, we showed that people’s prevalence judgments of a feature in a novel category were highly correlated with their judgments of the comparison category by directly providing participants the familiar category. We then proceeded to test our hypothesis using naturally occurring data, probing whether differences in participants’ demographic information may lead to differences in participants’ comparison categories and their estimations. Results from Experiments 3 and 4 revealed that across features people from different demographic groups make consistently different prevalence judgments for familiar and novel categories. Moreover, people’s prevalence judgments of novel categories were on average higher than their judgments of familiar categories given the same set of features. Following Tessler & Goodman (in press)'s work, We confirmed the highly context-dependent nature of generic language by showing that, upon hearing a novel generic, listeners' interpretation of the prevalence threshold of the given generic is highly flexible and sensitive to the prevalence threshold of comparison sets.

Recent theories of lexical knowledge proposes that multiple sources of information contribute to word meanings, and people's interpretation of words do not come context-free but heavily rely on other linguistic and non-linguistic cues, such as grammcatical structures and discourse constraints (Elman, 2009). The findings from our studies support this approach of understanding lexical knwoledge. By showing that people's interpretation of a novel category reflects their background knowledge of familiar categories (absent in the moment of discourse), we provide another piece of compelling evidence that the interpretation of a word is also dependent on listeners' prior experiences which occurred ahead of the moment of discourse. Moreover, because of this multi-factor composition of word meanings, the results from our studies further suggest a source of miscommunication between speakers and listeners due to differences in their prior experiences, despite their intentions to be cooperative. While in our studies we asked participants to explicitly estimate the prevalence ratings, this kind of distributional information is not always available or frequently exploited by conversational partners during a discourse. Speakers and listeners may not be aware of the asymmetrical intrepretations of the same statement due to their prior knowledge.

Although well-intentioned, misunderstandings arose due to differences in speaker's and listener's prior knowledge may still lead to serious consequences. Upon hearing an utterance by a speaker, listeners may generate interpretations that deviate far from speaker's intented meaning. Consider a situation where the speaker may produce a generic about a feature in a category with in mind a prevalence rating of 0.3. While sharing some common background that facilitated their cooperative conversation, due to the differences in their individual experiences, the listener may make an inference about the prevalence of the same feature in the given category to be 0.8, or vice versa. The direction of asymmetrical estimations between speaker and listener is highly depend on the prior distirbutional information a listener has access to and thus base her estimation on. When the category under discussion is a certain unfamiliar social group and some obscure feature, this asymmetry in listener's interpretation may further facilitate the transmission of stereotyping beliefs. In scenarios where listeners do not have access to other information to correct or update one's understanding, a listener may hold on to an erroneous belief and pass this inaccurate belief to others.

UNFINISHED Future directions: (1) probe how people choose the comparison categories upon hearing a generic. (2)There are more forms of generic that are studied in the linguistics literature, including noun phrases with indefinite articles and noun phrases that have definite articles. While in the current research only bare plural noun phrases was studied, more work should be done in the future to explore whether this of extend to all forms of genericity. (3) Another future direction for this work is to investigate how children would make estimation sample comparison sets from their prior experience from a developmental perspective. Generic languages are part of children's linguistic environment, but how are they interpretating the underlying distributional information?

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.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, *33*(4), 547–582.

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.

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*.