The Role of the QUD in Exhaustivity Inference from Generic Generalizations of Social Groups

Pratyusha R. Javangula
`pjavan@stanford.edu`

1 Introduction

As interlocutors communicate over the course of a conversation, their task is complicated by the fact that each of them intends meanings beyond what is said when they make an utterance. Therefore, comprehension is not merely the task of combining fixed lexical word meanings in ways specified by the syntax. Instead, listeners must make judgments about the intended meaning of an utterance that take into account the context in which the utterance was made, as well as more general information about "the way the world is". In order to develop a complete theory of language comprehension, it is imperative to investigate the nature of these judgments, more formally known as pragmatic inferences, and examine the linguistic factors that bring them about.

Much of the existing literature that seeks to answer questions about how pragmatic inferences arise focuses on a specific type of pragmatic inference known as *scalar implicature*. When making a scalar implicature, a listener infers "some, but not all" from "some" despite the fact that the lexical semantics of "some" taken out of context allows "some" to mean "all" (Carston, 1998). However, less research has been conducted into the factors influencing another type of pragmatic inference called exhaustivity inference. Exhaustivity inference is typically defined in the context of question answering and is defined as the inference made by the listener that the speaker's utterance alone constitutes an exhaustive list of answers to the question. For example, if a speaker answers the question "What did you have for dinner?" with "I had a salad," an exhaustive interpretation would be that the speaker *only* had a salad for dinner (van Rooij & Schulz, 2004).

Exhaustivity inferences are typically drawn from utterances that are answers to questions.

However, that question need not be an explicit utterance made over the course of a discourse. Rather, the question being answered could be the more implicit question under discussion. Roberts (2012) defined the question under discussion (QUD) as the question that interlocutors commit to finding an answer for over the course of a conversation. Furthermore, the QUD dictates what alternative meanings are available, given an utterance. For example, for the aforementioned utterance, "I had a salad for dinner," a QUD of, "What did you have for dinner?" proffers different alternative meanings than that of, "What did you do last night?"

Frank & Goodman (2012) present the Rational Speech Act framework as a means of formalizing the role of the QUD as well as a listener's prior beliefs in pragmatic inference. Under this framework, speakers are assumed to be informative, speakers and listened are assumed to be cooperative with respect to finding an answer for the QUD, and listeners' inferential mechanism is modeled as Bayesian inference. Applied to the generic generalization, "Black lives matter," the framework demonstrated that a pragmatic listener is more likely to draw an exhaustivity inference when they infer the QUD to be "Which lives matter?" rather than "Do Black lives matter?" (Degen, 2017). Further, Degen (2017) determined that a listener's prior beliefs influenced which QUD a listener would infer.

Here we present an empirical investigation into the findings of Degen (2017) that the QUD plays a role in exhaustivity inference from utterances that are generic generalizations of social groups. In this case, generic generalizations of social groups are utterances of the form, "`social group noun is predicate`." To conduct this investigation, we varied the context of generic generalizations via cover stories presented prior to the utter-

ance itself (Degen & Goodman, 2014). We then asked participants to assess the speaker's beliefs about the extent to which the social group mentioned in the utterance and its complement has the quality explicitly mentioned in the utterance. Finally, participants were asked to report their own beliefs about the extent to which the social groups previously presented have the qualities introduced in the previous utterances.

We hypothesized that participants would be more likely to draw an exhaustivity inference when presented with an utterance whose cover story induced a QUD of "Which `social group noun` is `predicate`?" compared to when presented with an utterance whose contextual cover story induced a QUD of "Is `social group noun predicate`?" This experiment was pre-registered on the Open Science framework and can be found here.

2 Method

2.1 Participants

We recruited 160 participants with US IP addresses and at least 95% of prior HITs approved on Amazon's Mechanical Turk platform (ages: 20-69; mean: 38.9). They were paid 1 dollar.

2.2 Materials

2.2.1 Cover Story Manipulation

The corpus of stimuli for the cover story manipulation phase comprised 32 combinations of a cover story and an utterance, where each utterance consisted of a social group noun paired with a predicate.

We selected four pairs of complementary social group nouns: *Democrats and Republicans, men and women, people in their 20's and people in their 50's* and *Europeans and Americans*. We also selected four positively valenced predicates: *competent, rational, polite* and *smart*. Utterances and cover stories containing *Democrats and Republicans* were always combined with the predicate *competent*, those containing *men and women* were always combined with the predicate *rational*, those containing *People in their 20's and People in their 50's* were always combined with the predicate *polite*, and those containing *Europeans and Americans* were always combined with the predicate *smart*. We decided on these specific fixed combinations because they reflect commonly-held stereotypes about each social group. An example utterance presented was the following:

(1) Democrats are competent.

For each of the four pairs of nouns, we created 8 stimuli. Each of the eight stimuli featured a cover story that evoked one of two QUD's. Each cover story was about a speaker in a meeting at their place of work with the larger group of people they work with, where a secondary actor announces that they will be choosing one or two subgroups for a project based on the extent to which the group(s) is/are `predicate`. Depending on if the secondary actor would be choosing one or two subgroups, the story induced an implicit QUD that made it relevant either that the speaker's utterance that "`social group noun` is `predicate`" should be pragmatically strengthened to "Only `explicitly mentioned group noun` is `predicate`" (*who?* condition) vs. "`explicitly mentioned noun` is `predicate`, and the extent to which `complementary noun` is `predicate` is not at issue" (*are?* condition). A cover story inducing the *who?* QUD follows:

(2) Michelle is a senator of a major political party and is a member of many legislative committees. Every legislative committee is composed of either all Democrats or all Republicans. The other day in a Senate session, the head of the Senate unveiled that he would be choosing an existing legislative committee to work on a new project based on their competence.

Now, we present an example cover story inducing the *are?* QUD:

(3) Michelle is a senator of a major political party and is a member of many legislative committees. Every legislative committee is composed of either all Democrats or all Republicans. The other day in a Senate session, the head of the Senate unveiled that he would be choosing two existing legislative committees to work on a new, highly classified project based on their competence, and that he had already chosen one team of Democrats.

The cover story in each stimulus had either a male or female speaker and an opposite-gender

secondary actor (e.g. *head of the Senate*). Finally, in each of the eight stimuli, the critical utterance explicitly mentions one of the two social group nouns in the pair.

2.2.2 Prior Beliefs Collection

The corpus of stimuli for the prior beliefs collection phase comprised eight questions of the format "How `predicate` do you find `social group noun`?" An example follows:

(4) How smart do you find Americans?

Stimuli were created by combining each of the eight nouns presented in the cover story manipulation phase with the predicate to which they were fixed.

2.3 Procedure

2.3.1 Cover Story Manipulation

Each participant was presented with a random one from each of the four sets of eight stimuli. Therefore, in total each participant responded to four stimuli, one for each pair of social group nouns. An important caveat to the random selection was that we presented two stimuli with cover stories inducing the *who?* QUD and two with cover stories inducing the *are?* QUD, such that QUD was a within-participants manipulation. Trial order was randomized between-participants.

First, participants were presented with the cover story and asked to read it in its entirety. Then, to ensure that participants paid attention to the story, they had to answer the following comprehension question: *Has the* `secondary actor` *chosen at least one* `subgroup`? They were only allowed to proceed once they correctly answered the question. The correct answers were either *yes* if the cover story induced the *are?* QUD or *no* if the cover story induced the *who?* QUD.

They were then asked *How* `predicate` *are* `social group noun`, *according to* `speaker`? They then adjusted two sliding scales with endpoints labeled extremely `opposite of predicate` and extremely `predicate` to indicate their judgments of the speaker's beliefs of both the group noun in the utterance and its complement. Slider values were recorded as falling between 0 ("extremely `opposite of predicate`") and 1 ("extremely `predicate`").

For our purposes, the opposite of *smart* is *stupid*, the opposite of *rational* is *emotional*, the

David is a student at an international high school and is a member of many study groups. Every study group in his school is composed of either all American or all European students. The other day in an all-school assembly, the principal unveiled that she would be choosing an existing study group to receive a prestigious scholarship based on their intelligence. David raised his hand to speak in the assembly and said,

Americans are smart.

How smart are Americans, according to David?

Extremely stupid   Extremely smart

How smart are Europeans, according to David?

Extremely stupid   Extremely smart

Continue

Figure 1: Distribution of gender over samples in training data set.

opposite of *polite* is *rude*, and the opposite of *competent* is *incompetent*. Figure 1 presents a screenshot of a standard cover story manipulation trial.

If an exhaustivity inference is drawn, meaning that the speaker is interpreted as saying that only the critical noun from the utterance is `predicate`, participants should judge that group noun as having a greater value for `predicate` than the complementary group noun. If instead they do not draw an exhaustivity inference, the difference in their judgments for the extent to which each group is `predicate` should be smaller.

2.3.2 Prior Beliefs Collection

Each participant was presented with each of the eight stimuli created for the prior beliefs collection phase. Trial order was randomized between-participants.

They asked *How* `predicate` *do you find* `social group noun`? They adjusted two sliding scales with endpoints labeled extremely `opposite of predicate` and extremely `predicate` to indicate their responses for both the group noun in the utterance and its complement. Slider values were coded in the same way as for the cover story manipulation phase.

3 Results

Mean slider response of target social group noun (mentioned directly in the utterance) and the other social group noun (its complement) as a function of the context-induced QUD are shown in Figure 2. A visual assessment of the difference in mean response between *who?* and *are?* conditions seems to suggest that in line with our predictions, there is a marginally smaller difference between mean response for the target and other nouns in the *are?* condition compared to the *who?* con-
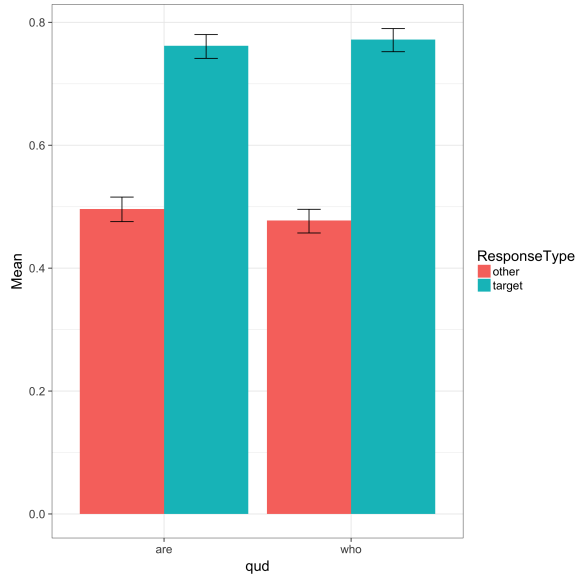
Figure 2: Mean response to as a function of QUD to target and other social group noun.

dition. However, this difference does not suggest a significant effect of QUD manipulation. There does appear to be a significant difference between mean response for the target social group noun and the other social group noun.
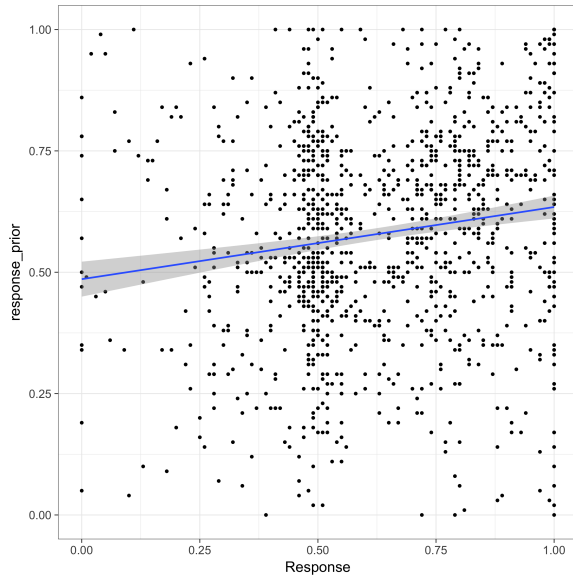

Figure 3: Response to noun (target and other) as a function of prior belief.

We also investigated whether participants' assessments of the speaker's beliefs of the target social group noun and other social group noun were related in any way to their own beliefs. Figure 3 presents participant prior beliefs as a function of

their judgments of the speaker's beliefs, with a linear smoother and 95% confidence intervals.

Examining the slope of the linear smoother shows a slightly positive correlation between a participant's judgment of the speaker's beliefs about a social group noun and a participant's own belief about that social group noun. To verify the conclusions made from inspection of Figure 3, we conducted a Pearson correlation between judgments of the speaker's beliefs and found that they were weakly but significantly correlated, $r(1200) = 0.1557$, $p < .001$.

After observing this significant correlation, in addition to investigating the effect of QUD manipulation on participant judgments of speaker beliefs, we sought to address the additional question of the role that prior beliefs play in judgments of speaker beliefs, if any.

To address these questions, we conducted a Bayesian generalized multivariate non-linear regression using the `brms` package (Büerkner, 2017). Our purpose for using Bayesian methods was that it would allow us to fit a model with a full random effects structure where mixed-effects linear models instantiated using `lme4` would not.

The model predicted the log odds of slider response representing a participant's judgment of the speaker's beliefs about the extent to which `social group noun` is `predicate` from fixed effects of the QUD, response type (whether the response given is in reference to the critical noun versus the other noun), their interaction, and the participant's prior beliefs. In order to avoid problems related to collinearity, we centered all predictor variables. We included the maximal random effects structure justified by the design: random by-participant and by-item slopes for the interaction between QUD and response type and the prior belief (reflecting the possibility that the effects of these variables could vary by participant and by item). Four chains converged after 2000 iterations each (warmup = 1000, $\hat{R} = 1$ for all estimated parameters). T

We report 95% confidence intervals and the posterior probabilities $P(\beta > 0)$ that coefficients for QUD, response type, their interaction, and prior beliefs are greater than zero. We also provide 95% confidence intervals (CI) of the posterior distributions for each coefficient. Finally, if zero is not included in the CI for the posterior distribution of a variable's coefficient, we have strong reason to be-

lieve it has a significant effect on slider response. So too, if $P(\beta > 0)$ is close to zero or one.
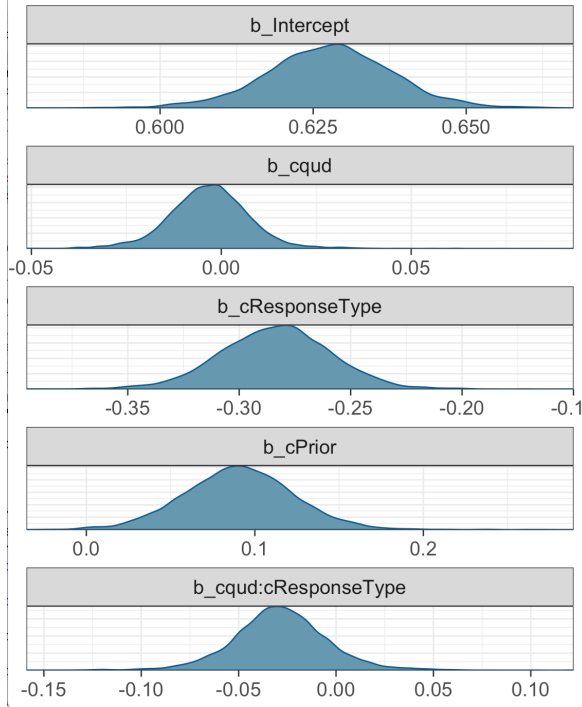

Figure 4: Posterior distributions for all of model predictors.

Table 1 presents posterior mean β, 95% CI, and $P(\beta > 0)$ for each of the predictors to the model, and Figure 4 presents the posterior distributions of the coefficients of each predictor. From Table 1 and Figure 4 together it is clear that there is a significant effect of response type and prior, but no significant effect of QUD and no significant interaction. These results align with the intuitions obtained from examining the data visualizations. Nevertheless, we find that our original predictions that QUD would have a significant effect on participant judgments of speaker beliefs were not borne out. We interpret these results, reason about their origins, and propose future directions in the next section.

4 Discussion

The purpose of the present research was to assess whether there is empirical evidence to support the finding from previous work using computational tools to model pragmatic inference that inferred QUD has an effect on exhaustivity inference in response to generic generalizations of social groups. This aim was carried out by presenting participants with a series of utterances that were generic generalizations of different social groups, accompanied by cover stories intended to spur the participant to infer one of two QUDs. We then recorded participants' judgments of the speaker's beliefs about the extent to which the social group implicated in the utterance and its unmentioned complement embody the predicate of the utterance. Finally, we collected participants' prior beliefs about the extent to which each of the previously presented social groups embodies the predicate it was fixed to in the previous phase.

Prior to conducting this research, we operationalized an exhaustivity inference as a large difference between a participant judgment of a speaker's beliefs about the social group explicitly mentioned in the utterance and its unmentioned complementary social group. We hypothesized that when presented with a cover story inducing the *who?* QUD, participants would be more likely to draw an exhaustivity inference than when they were presented with a cover story inducing the *are?* QUD.

The empirical pattern of obtained results is as follows. We observed effects of response type and priors on participant judgment of speaker beliefs of the extent to which a certain social group embodies a given predicate. On the other hand, we did not observe an effect of QUD or of the interaction between QUD and response type.

Although not hypothesized, the observed main effect of response type validates the study design. Since participants gave significantly different responses for the target social group noun compared to the unmentioned other social group noun, we can conclude that our cover story and utterance presentation appropriately informed participants that the the extent to which the two groups embodied the predicate are different.

Examining the pattern of results in Figure 2, however, it is surprising that in the *are?* condition, there is such a large difference between the responses to target and other nouns. Since the *are?* QUD is intended to prime participants to believe the complementary social noun is irrelevant, and there is no utterance explicitly mentioning the complementary social noun, it is surprising that participants gave low judgments to the other noun, on average. We suspect this phenomenon arose because all participants were not uniformly inferring the correct QUD across all trials. In future work, we aim to employ the approach taken by Degen &

Variable	β	95% CI	$P(\beta) > 0$
QUD	0.00	[-0.03,0.02]	0.2331
response type	-0.28	[-0.33,-0.24]	0
QUD * response type	-0.03	[-0.08,0.02]	0.1015
prior beliefs	0.09	[0.03,0.16]	0.9963

Table 1: Posterior mean β, 95% confidence interval, and $P(\beta) > 0$ for all fixed effects and interactions.

Goodman (2014) and explicitly include a comprehension question to ensure correct interpretation of the QUD from the cover story.

Similarly, the observed main effect of prior beliefs on slider response was not hypothesized. We find it reasonable that a participant's own prior beliefs would have an effect on their judgments of a speaker's beliefs. Furthermore, these results reaffirm the findings of Degen (2017) that prior beliefs have a role to play in pragmatic inference from generic generalizations of social groups.

Most importantly, there was no observed effect of QUD, which contradicts the hypotheses formulated prior to conducting the experiment. In discussing why we believe these results may have arisen, we will address a major limitation of this work.

Aside from the aforementioned limitation that participants may not have inferred the correct QUD for the condition they were assigned to in a particular trial, we believe these results came about due to the nature of the interaction between the social group noun and predicate in certain utterances.

Figure 5 presents mean response as a function of QUD to the target and complementary noun, separated by pair. From Figure 5 we can see that the hypothesized effect of QUD occurs when participants responded to utterances about men or women, but not for utterances about any other social group nouns.

This item-specific effect may have arisen because each of the generic generalizations of social groups may have been assigned a different meaning, regardless of the cover story that accompanied it. (Tessler & Goodman, 2016) found that generic statements like "Dogs bark" and "Robins are female" take into account diverse sets of prior beliefs about different properties for each statement and thus are interpreted vastly differently in and out of context.

The question of how meaning is assigned to statements of this nature about social groups is rel-
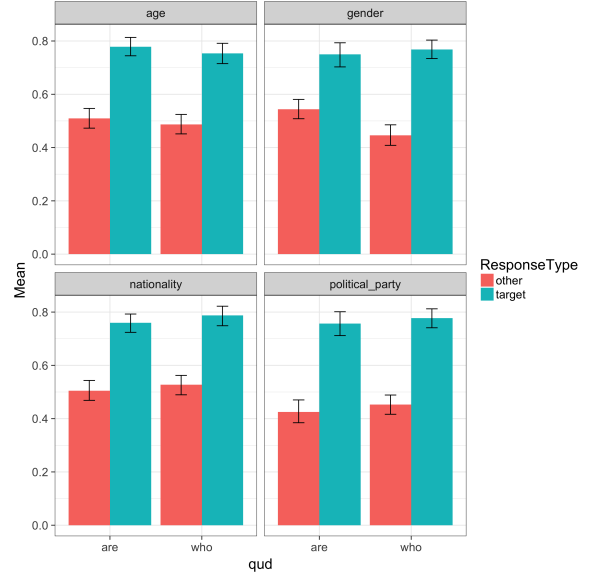

Figure 5: Mean response to as a function of QUD to target and other social group noun.

atively unexplored. Thus, in order to meaningfully explore the role of the QUD in exhaustivity inference of these types of utterance, it will be necessary to extend the findings of (Tessler & Goodman, 2016) to utterances like those that we presented as stimuli in the current work.

In light of the vastly different results for different items, future work must prioritize an investigation of how existing pragmatic theories of generic language apply to generic statements about social group nouns, as well as a separate investigation of the role of the QUD in exhaustivity inference of non-generic statements like "I had a salad for dinner."

5 Conclusion

We have shown that overall, presentation of different cover stories meant to evoke different QUD's prior to an utterance plays no role in exhaustivity inference of generic generalizations of social groups. However, we were able to show that a participant's prior beliefs, as well as the fact of whether a participant gave a response about an ex-

plicitly mentioned or complementary social group noun, can greatly affect a participant's judgment of speaker beliefs about both the noun in the utterance and the complementary noun.

Acknowledgments

References

Büerkner, P. C. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, *80*(1), 1–28. doi: 10.18637/jss.v080.i01

Carston, R. (1998). Informativeness, relevance and scalar implicature. In S. Carston R. & Uchida (Ed.), *Relevance threory: Applications and implications.* (p. 176-239). Amsterdam: John Benjamins.

Degen, J. (2017). *Black lives matter: the role of prior beliefs and the qud in (preventing) exhaustivity inferences.* (Formal Meaning and Social Pragmatics Workshop)

Degen, J., & Goodman, N. (2014). Lost your marbles? the puzzle of dependent measures in experimental pragmatics. In *Proceedings of the 36th annual meeting of the cognitive science society* (pp. 397–402). Cognitive Science Society.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998. Retrieved from `http://science.sciencemag.org/content/336/6084/998` doi: 10.1126/science.1218633

Roberts, C. (2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, *5*(6), 1–69.

Tessler, M. H., & Goodman, N. (2016). *A pragmatic theory of generic language* (Unpublished doctoral dissertation). Stanford University.

van Rooij, R., & Schulz, K. (2004, Mar 01). Exhaustive interpretation of complex sentences. *Journal of Logic, Language and Information*, *13*(4), 491–519. Retrieved from `https://doi.org/10.1007/s10849-004-2118-6` doi: 10.1007/s10849-004-2118-6