Warm (for winter): Comparison class understanding in vague language

Michael Henry Tessler[1], **Michael Lopez-Brau**[2], and **Noah D. Goodman**[1]
mtessler@stanford.edu, lopez_mic@knights.ucf.edu, ngoodman@stanford.edu
[1]Department of Psychology, Stanford University, [2]Department of Electrical & Computer Engineering, University of Central Florida

Abstract

Speakers often refer to context only implicitly when using language. The utterance "it's warm outside" could mean it's warm relative to other days of the year or just relative to the current season (e.g., it's warm for winter). *Warm* vaguely conveys that the temperature is high relative to some contextual *comparison class*, but little is known about how a listener decides upon such a standard of comparison. Here, we propose that the resolution of a comparison class in context is a pragmatic inference driven by world knowledge and listeners' internal models of speech production. We introduce a Rational Speech Act model and derive two novel predictions from it, which we validate using a paraphrase experiment to measure listeners' beliefs about the likely comparison class used by a speaker. Our model makes quantitative predictions given prior world knowledge for the domains in question. We triangulate this knowledge by a second language task in the same domains, using Bayesian data analysis to infer priors from both data sets.

Keywords: comparison class; pragmatics; Rational Speech Act; Bayesian cognitive model; Bayesian data analysis

If it's 75 °F (24 °C) outside, you could say "it's warm." If it's 60 °F (16 °C), you might not consider it warm. Unless it's January; it could be warm for January. *Warm* is relative, and its felicity depends upon what the speaker uses as a basis of comparison—the *comparison class* (e.g., other days of the year or other days in January). Comparison classes are necessary for understanding adjectives and, in fact, any part of language whose meaning must be pragmatically reconstructed from context, including vague quantifiers (e.g., "He ate a lot of burgers."; Scholler & Franke, 2015) and generic language (e.g., "Dogs are friendly"; Tessler & Goodman, 2016a). The challenge for listeners is that the comparison class often goes unsaid (e.g., in "it's warm outside").

The existence of comparison classes for understanding vague language is uncontroversial (Bale, 2011; Solt, 2009). 4-year-olds categorize novel creatures (*pimwits*) as either "tall" or "short" depending on the distribution of heights of *pimwits* and not the heights of creatures that are not called *pimwits*, suggesting the comparison class here is *other pimwits* (Barner & Snedeker, 2008). Adult judgments of the felicity for adjectives like "dark" or "tall" similarly depend upon fine-grained details of the statistics of the comparison class (Qing & Franke, 2014b; Schmidt, Goodman, Barner, & Tenenbaum, 2009; Solt & Gotzner, 2012).

It is not well understood, however, how a comparison class becomes contextually salient. Any particular object of discourse can be conceptualized or categorized in multiple ways, giving rise to multiple possible comparison classes. A day in January is also a day of the year; if it's warm, it could be *warm for Winter* or *warm for the year*. We propose that listeners actively combine category knowledge with knowledge about what classes are likely to be talked about to infer the implicit comparison class used by the speaker. We introduce a minimal extension to the Rational Speech Act (RSA) model for gradable adjectives (Lassiter & Goodman, 2013) to allow it to flexibly reason about the likely comparison class.

From this model, we derive two novel qualitative predictions. For illustration, consider temperature. In Winter, the positive adjective phrase "it's warm" should signal a specific class (i.e., "it's warm relative to days in Winter") more so than the negative phrase "it's cold," which could signal a specific or general class (e.g., "it's warm relative to days of the year). The opposite should be true in Summer, where *negative* forms should be more likely to signal the specific class (i.e., cold for Summer). This is driven by the *a priori* probability that the adjective could apply to the specific class e.g., the chance that it's warm in Winter (Prediction 1). In addition, regardless of the season, a listener should prefer"warm" or "cold" to be relative to the current season, a more specific reference class. All else being equal, classes that are more specific will have relatively lower variance, which results in a more specific meaning for a vague utterance (Prediction 2). We test these predictions in an experiment designed to elicit the comparison class using a paraphrase dependent measure (Expt. 1).

The RSA model makes quantitative predictions given fine-grained details of the relevant prior knowledge: the distribution of temperatures for days in winter, days in the whole year, and so on. Background knowledge of this sort has previously been measured by having participants estimate quantities or give likelihood judgments directly (Franke et al., 2016). We pursue a different methodology. Because our model captures a productive fragment of natural language, it makes predictions about simpler linguistic tasks (based on the same background knowledge). In Expt. 2 we collect judgments about sentences like "This [winter day] is cold relative to other days of the year," in which domain knowledge is needed but the comparison class is explicit. We then use a Bayesian data analysis model to infer the parameters of the background knowledge from linguistic judgments. By harnessing the productive nature of language into experiment design, we are able to learn jointly about interlocutors' latent knowledge and language use.

Understanding comparison classes

Adjectives like *warm* and *cold* are vague descriptions of an underlying quantitative scale (e.g., temperature). The vagueness and context-sensitivity of these adjectival utterances can be modeled using threshold semantics ($[\![u]\!] = x > \theta$, for utterance u, scalar degree x, and threshold θ), where the threshold comes from an uninformed prior distribution and is inferred in context via pragmatic reasoning (Lassiter & Goodman, 2013; see also Qing & Franke, 2014a):

$$L_1(x,\theta \mid u) \propto S_1(u \mid x,\theta) \cdot P_c(x) \cdot P(\theta) \quad (1)$$

$$S_1(u \mid x,\theta) \propto \exp(\alpha_1 \cdot \ln L_0(x \mid u,\theta)) \quad (2)$$

$$L_0(x \mid u,\theta) \propto \delta_{[\![u]\!](x,\theta)} \cdot P_c(x) \quad (3)$$

This is a Rational Speech Act (RSA) model, a recursive Bayesian model where speaker S and listener L coordinate on an intended meaning (for a review, see Goodman & Frank, 2016). In this framework, the pragmatic listener L_1 tries to resolve the state of the world x (e.g., the temperature) from the utterance she heard u (e.g., "it's warm"). She imagines the utterance coming from an approximately rational Bayesian speaker S_1 trying to inform a naive listener L_0, who in turn updates her prior beliefs $P(x)$ via an utterance's literal meaning $[\![u]\!](x)$. Lassiter & Goodman (2013) introduced into RSA uncertainty over a semantic variable: the truth-functional threshold θ (Eq. 1). The uncertainty over θ (e.g., the point at which something is *warm*) interacts with the prior distribution over possible states of the world $P_c(x)$ (e.g., possible temperatures) to resolve the meaning of the adjective in context. The prior distribution over world-states is always relative to some comparison class c (Eqs. 1 & 3) but where does the comparison class come from?

When a listener hears only that "it's warm outside" without an explicit comparison class (e.g., "warm for the season"), we imagine the listener must infer the comparison class by thinking about what would best complete the sentence. She does this using her world knowledge of what worlds are plausible given different comparison classes $P(x \mid c)$, what comparison classes are likely to be talked about $P(c)$, and what a rational speaker would say in given a world and comparison class $S_1(u \mid x,c,\theta)$ (Eq. 4). As a first test of this idea, we consider an idealized case where the comparison class can be either a relatively specific (subordinate) or relatively general (superordinate) categorization (e.g., warm relative to days in Winter or relative to days of the year). The listener is aware that the target entity is in the subordinate class (e.g., aware that it is winter) and draws likely values of the degree (e.g., temperature) from the subordinate class prior $P(x \mid c_{sub})$. With these assumptions, the model becomes:

$$L_1(x,c,\theta \mid u) \propto S_1(u \mid x,c,\theta) \cdot P(x \mid c_{sub}) \cdot P(c) \cdot P(\theta) \quad (4)$$

$$S_1(u \mid x,c,\theta) \propto \exp(\alpha_1 \cdot \ln L_0(x \mid u,c,\theta)) \quad (5)$$

$$L_0(x \mid u,c,\theta) \propto \delta_{[\![u]\!](x,\theta)} \cdot P(x \mid c) \quad (6)$$

We are interested in the behavior of the model with the underspecified utterance (e.g., "It's warm"), and we assume the speaker has two alternative utterances in which the comparison class is explicit (e.g., "It's warm relative to other days in Winter." and "It's warm relative to other days of the year.").

The quantitative predictions of this model depend on the details of the listener's knowledge of the subordinate and superordinate categories: $P(x \mid c_{sub})$ and $P(x \mid c_{super})$, as well as the prior distribution on comparison classes $P(c)$ in Eq. 4.

Comparison class prior $P(c)$ reflects listeners' expectations of what classes are likely to be discussed. As a proxy for comparison class usage frequency, we use empirical frequency $\hat{f}$ estimated from the Google WebGram corpus[1], and scale it by a free parameter β such that $P(c) \propto \exp(\beta \cdot \log \hat{f})$.

Degree priors (world knowledge) Only the relative values for $P(x \mid c_{sub})$ and $P(x \mid c_{super})$ affect model predictions. Hence we fix each superordinate distribution to be a standard normal distribution $P(x \mid c_{super}) = \mathcal{N}(0,1)$ and the subordinate priors to also be Gaussian distributions $P(x \mid c_{sub}) = \mathcal{N}(\mu_{sub}, \sigma_{sub})$; the subordinate priors thus have standardized units. We will eventually infer the parameters of the subordinate priors from experimental data.

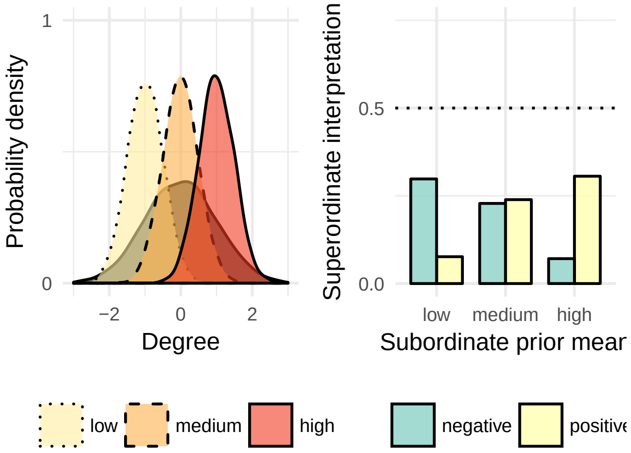

Figure 1: Left: Three hypothetical prior distributions over subordinate comparison class (fixing the superordinate comparison class to be a unit-normal distribution, in grey). Right: Predicted listener inference for the probability of an intended superordinate comparison class. The comparison class prior does not favor either the subordinate or superordinate class.

Qualitative model predictions For purposes of illustration, assume for the moment that each class is equally likely *a priori*: $P(c) = 0.5$. Figure 1 (left) illustrates superordinate and possible subordinate priors; e.g., temperatures over the whole year, in Winter (low), Fall (medium), and Summer (high), in a part of the world that exhibits seasons (e.g., Maryland, USA). We see that regardless of the adjective polarity (e.g. warm vs. cold) or the mean of the subordinate prior (low, medium, high), the pragmatic listener (Eq. 4) prefers the more specific comparison class (i.e., the subordinate class; Figure 1, right; all bars below baseline). We also see that

[1] Corpus accessed via `http://corpora.linguistik.uni-erlangen.de/demos/cgi-bin/Web1T5/Web1T5_freq.perl`. Due to the idiosyncracies of our categories and potential polysemy in some of the words (Table 1), we made the following substitutions when querying the database for emprical frequency: produce → "fruits and vegetables"; things you watch online → "online videos"; days in {season} → "{season} days"; dishwashers → "dishwashing machines"; videos of cute animals → "animal videos".

Scale (adjectives)	Subordinate classes	Superordinate
Height (tall, short)	(professional) gymnast, soccer player, basketball player	people
Price (expensive, cheap)	bottle opener, toaster, dishwasher	kitchen appliances
Temperature (warm, cold)	Winter, Fall, Summer (day in Maryland)	days in the year
Time (long, short)	video of a cute animal, music video, movie	things you watch online
Weight (heavy, light)	grape, apple, watermelon	produce

Table 1: Items used in Experiments 1 and 2. Subordinate categories were designed to fall near the low end, high end, and somewhere in the middle of the degree scale

during Winter, when the pragmatic listener hears "it's warm" (positive adjective), she thinks it's more likely that it's warm relative to days in Winter than when she hears "it's cold" (Figure 1, right, left-most bars). The opposite is true in Summer. In sum, we see two predictions: The pragmatic listener overall prefers subordinate comparison classes, though the extent of this preference is modulated by the *a priori* probability that the adjective is true of the superordinate category (which depends on its polarity). We will test these two predictions in our first experiment.

Reducing uncertainty with productive language use As we've described above, the relevant prior knowledge yields two free parameters per subordinate domain. Even if the parameters could be estimated reliably from our target (comparison class) expriment, it's hard to know if the parameters we're uncovering are the true priors for participants' knowledge, or just extra parameters used to fit the data. However, the RSA model is a productive language model: We can use it to predict related language data that rely on the same prior knowledge. In Expt. 2, we gather judgments about sentences with vague adjectives but explicit comparison classes: whether or not an adjective would be true relative to the superordinate category of a subordinate member (e.g., it is a winter day, is it warm relative to other days of the year?).

For this data, we remove comparison class uncertainty by setting $P(c_{super}) = 1$, since the sentences provide explicit comparison classes. We then model sentence endorsement using a pragmatic speaker (following Qing & Franke, 2014a; Tessler & Goodman, 2016a, 2016b):

$$S_2(u \mid c_{sub}) \propto \exp\left(\alpha_2 \cdot \mathbb{E}_{x \sim P_{c_{sub}}} \ln L_1(x \mid u)\right) \qquad (7)$$

Note that $L_1(x \mid u)$ is defined from Eq. 4 by marginalization.

Eqs. 4 and 7 define models for the data we will gather from experiments 1 and 2, and depend on the same background knowledge. We can thus use data from both experiments to reconstruct the prior knowledge and generate predictions for the two data sets. Experimental paradigms, computational models, preregistration report, and data for this paper can be found at `https://mhtess.github.io`.

Behavioral experiments

The materials and much of the design of the two experiments are shared. Participants were recruited from Amazon's Mechanical Turk and were restricted to those with U.S. IP addresses with at least a 95% work approval rating. Each experiment took about 5 minutes and participants were compensated $0.50 for their work.

Materials We used positive- and negative-form gradable adjectives describing five scales (Table 1). Each scale was paired with a superordinate category, and for each superordinate category, we used three subordinate categories that aimed to be situated near the high-end, low-end, and intermediate part of the degree scale (as in Figure 1). This resulted in 30 unique events ({3 subordinate categories} x {5 scales} x {2 adjective forms}). Each participant saw 15 trials: one for each subordinate category paired with either the positive or negative form of its corresponding adjective. Participants never judged the same subordinate category for both adjective forms (e.g., cold and warm Winter days) and back-to-back trials involved different scales to avoid fatigue.

Experiment 1: Comparison class inference

In this experiment, we gather human judgments of comparison classes in ambiguous contexts, testing the two predictions described in **Qualitative Model Predictions**.

Participants and procedure We recruited 264 participants and 2 were excluded for failing an attention check. On each trial, participants were given a context sentence to introduce the subordinate category (e.g., *Tanya lives in Maryland and steps outside in Winter.*). This was followed by an adjective sentence, which predicated either a positive- or negative-form gradable adjective over the item (e.g., *Tanya says to her friend, "It's warm."*). Participants were asked "What do you think Tanya meant?" and given a two-alternative forced-choice to rephrase the adjective sentence with either an explicit subordinate or superordinate comparison class:

> {She / He / It} is ADJECTIVE (e.g., warm) relative to other SUBORDINATES (e.g., *days in Winter*) or SUPERORDINATES (e.g., *days of the year*)

In addition to all of the above design parameters, half of our participants completed trials where an additional sentence introduced the superordinate category at the beginning (e.g., *Tanya lives in Maryland and checks the weather every day.*), with the intention of making the superordinate paraphrase more likely.

Results We observe no systematic differences between participants' responses when the superordinate category was pre-

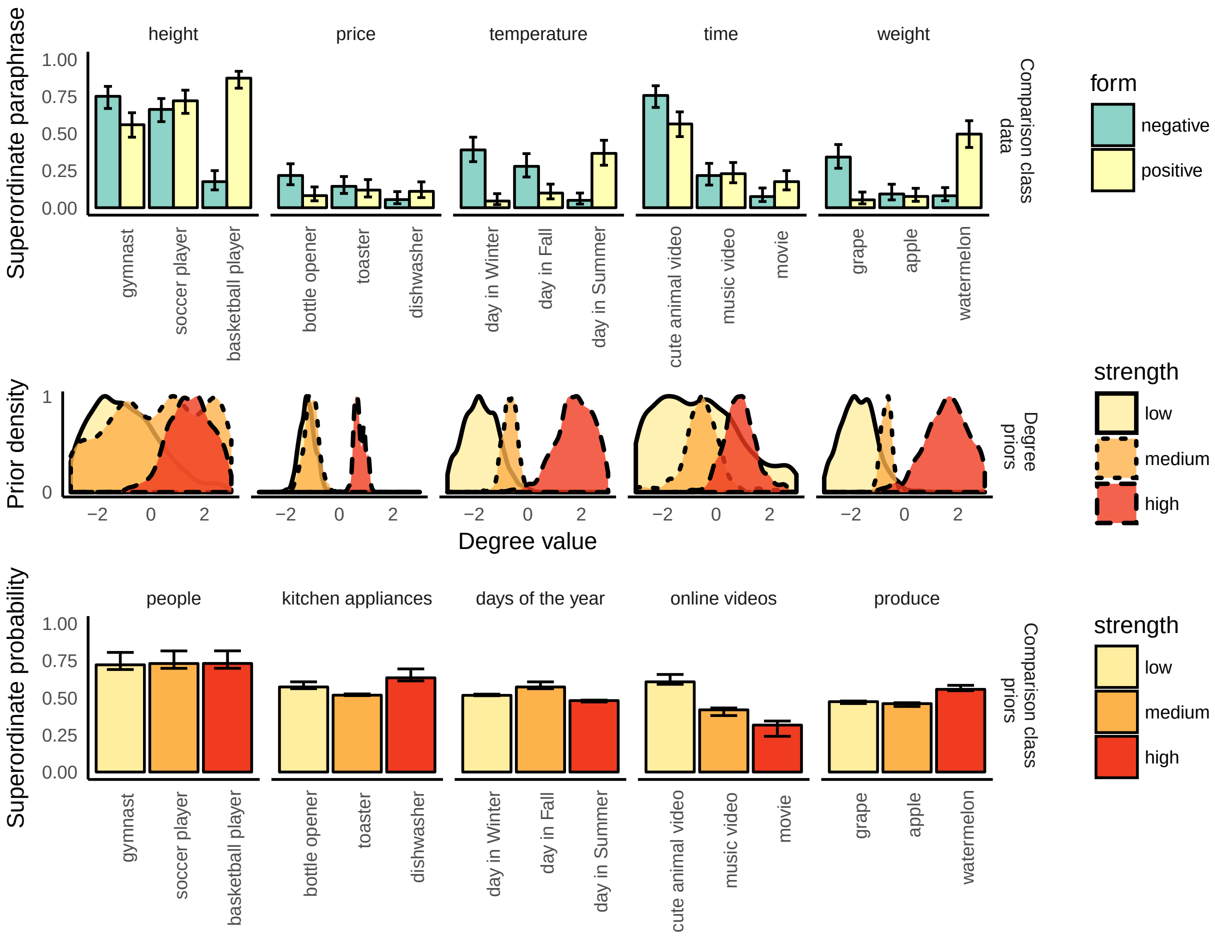

Figure 2: Empirical data, model inferred world priors, and empirically measured comparison class priors. Top: Experiment 1 results. Comparison class judgments in terms of proportion judgments in favor of superordinate comparison class. Error bars correspond to 95% Bayesian credible intervals. Middle: Inferred prior distributions of world knowledge used in Experiments 1 and 2. Bottom: Inferred prior probability of the superordinate comparison class based on Google WebGram frequencies. Error bars are derived from the 95% credible interval on the β scale parameter.

viously mentioned in the context and those when it was not; thus we collapse across these two conditions for subsequent analyses. Figure 2 (top) shows the proportion of participants choosing the superordinate paraphrase for each item, revealing considerable variability both within- and across- scales. The predicted effects are visually apparent within each scale.

Our predictions are confirmed using a generalized linear mixed effects model with main effects of adjective form (positive vs. negative) and the *a priori* judgment by the first author of whether the sub-category was expected to be low or high on the degree scale, and of critical theoretical interest, the interaction between these two variables. In addition, we included by-participant random effects of intercept and by-subordinate category random effects of intercept and iteraction between form and strength[2]. Confirming our two qualitative model predictions, there was an interaction between form and strength ($\beta = 3.75$; $SE = 0.58$; $z = 6.49$) and there was an overall preference for subordinate category paraphrases ($\beta = -1.21$; $SE = 0.37$; $z = -3.27$). The effects of form and strength were not significant.

We then test the simple effects. For items on the low end of the scale (e.g., temperature in Winter), positive form adjectives are significantly more likely to lead *away* from superordinate comparison classes ($\beta = -1.41$; $SE = 0.15$; $z = -9.43$), while the opposite is true for items high on the scale (e.g., Summer days; $\beta = 2.5$; $SE = 0.19$; $z = 13.15$).

Experiment 2: Adjective endorsement

In this experiment, we use a two-alternative forced choice paradigm to collect participants' adjective endorsements using knowledge that would be relevant for Expt. 1.

Participants and procedure We recruited 100 participants and 5 were excluded for failing an attention check. On each trial, participants were given a sentence introducing the subordinate category (e.g., *Alicia lives in Maryland and steps*

[2] This was the maximal mixed-effects structure that converged.

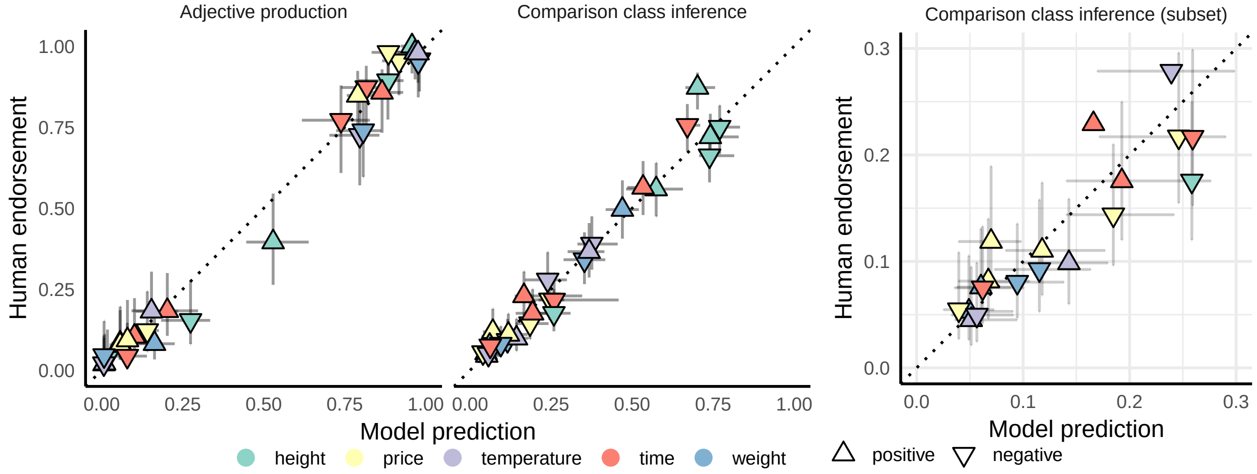

Figure 3: Human endorsement of comparison class paraphrases (middle) and adjective sentences (left) as a function of listener model L_1 and speaker model S_2 predictions, respectively. The right facet displays a subset of the paraphrase data (Expt. 1) to reveal good quantitative fit even in a small dynamic range. All error bars correspond to 95% Bayesian credible intervals.

outside in Winter.). This was followed by a question asking if the participant would endorse an adjective explicitly relative to the superordinate category (e.g., *Do you think the day in Winter would be warm relative to other days of the year?*).

Results The results of this experiment were as expected and can be seen roughly in Figure 3 (left, y-axis). We see that the endorsement of adjectival phrases in these domains is markedly more categorical than the comparison class inference task (compare vertical spread of left and middle facets).

Full model analysis and results

The full RSA model has a number of parameters; we now describe the Bayesian data analysis model used to treat these parameters and generate predictions. The comparison class prior uses scaling parameter on the empirical frequency $\hat{f}$, which we give the following prior: $\beta \sim \text{Uniform}(0,3)$. We put the same priors over the parameters of each subordinate Gaussian: $\mu \sim \text{Uniform}(-3,3)$, $\sigma \sim \text{Uniform}(0,5)$, since they use standardized units.

The full model has three additional parameters not of direct theoretical interest: the speaker optimality parameters α_i^{expt}, which can vary across the two tasks. Expt. 1 uses the pragmatic listener L_1 model (Eq. 4), which has one speaker optimality: α_1^1. Expt. 2 uses the pragmatic speaker S_2 model (Eq. 7), which has two speaker optimality parameters: $\{\alpha_1^2, \alpha_2^2\}$. We use priors consistent with the previous literature: $\alpha_1 \sim \text{Uniform}(0,20)$, $\alpha_2 \sim \text{Uniform}(0,5)$

Results

We implemented the RSA and Bayesian data analysis models in the probabilistic programming language WebPPL (Goodman & Stuhlmuller, 2014). To learn about the credible values of the parameters and the predictions of the model, we used an incrementalized version of MCMC (Ritchie, Stuhlmuller, & Goodman, 2016), collecting 2 independent chains of 75,000 iterations (removing the first 25,000 for burn-in).

The full model's posterior over the RSA and data-analytic parameters were consistent with results in prior literature and intuition. The maximum a-posteriori (MAP) estimate and 95% highest probability density (HPD) intervals for model parameters specific to the L_1 model used for Expt. 1 were $\alpha_1^1 = 1.59[1.12, 2.47]$, $\beta = 0.13[0.11, 0.19]$. Model parameters specific to the S_2 model used for Expt. 2: $\alpha_1^2 = 3.53[0.57, 13.23]$, $\alpha_2^2 = 3.24[2.61, 3.84]$.

The inferred values and shapes of the subordinate class knowledge used in these tasks is also consistent with intuition, as can be see in Figure 2 (middle). The items judged *a priori* to be low the on scale (yellow) tend to be lower than those judged to be in the middle of the scale (orange) and high on the scale (red).

Finally, the full model's posterior predictive distribution does an excellent job at capturing the variability in responses for Expt. 1: $r^2(30) = 0.965$, and Expt. 2: $r^2(30) = 0.985$ (Figure 3). Because of the overall preference for the subordinate comparison class, many of the data points are distributed below 0.5. Even for these fine-grained inferences, the model does a good job at explaining human participants' behavior (Figure 3 right).

Discussion

The words we say are often too vague to have a single, precise meaning, and only make sense in context. Context, however, can also be underspecified, as there are many possible dimensions or categories that the speaker might be implicitly referring to or comparing against. Here, we investigate the flexibility in the reference class against which an entity can be implicitly compared. We introduced a minimal extension to an adjective interpretation RSA model in order to flexibly reason about the likely comparison class. This model made two novel predictions about the comparison class listeners should prioritize. It also made quantitative predictions about how background knowledge about the degree scale should inform this inference. The quantitative and both of the qualitative predictions from the model were borne out in our first experiment. To our knowledge, this is the first experiment to demonstrate how reference classes for adjective interpretation can adjust based on world knowledge.

In our modeling work, we had to specify a prior distribution over the two comparison class alternatives $P(c)$, used in Expt. 1. There are at least two (non-mutually exclusive) ways of interpreting this prior distribution. The prior may reflect basic level effects in categorization (Rosch & Mervis, 1975) or the heterogeneity of the property within different classes. For instance, the distribution over prices for kitchen appliances may be more heterogeneous than the distribution of heights for people, making "kitchen appliances" a relatively poorer reference class for the price of a specific kitchen appliance than "people" is for the height of a specific person. A more heterogeneous class would make a vague utterance less informative. Alternatively, the adjective phrase "It's warm outside" may in fact be an incomplete sentence, in a way analogous to sentence fragments studied in noisy-channel models of production and comprehension (Bergen & Goodman, 2015). If this is so, we would expect $P(c)$ to be correlated with factors relevant for speech production, e.g., the frequency of the comparison class in a corpus. Future work should attempt to disentangle these factors to construct a more complete theory of the comparison class prior.

The second contribution of this paper is a novel data-analytic approach, where prior knowledge used in the Bayesian language model is reconstructed from converging evidence gathered from experiments that use similar language about the same domains. In previous work, we have attempted to measure prior knowledge by decomposing what would be an implicitly multilayered estimation question into multiple simpler questions, and then using a Bayesian data analytic model to reconstruct the prior knowledge (Tessler & Goodman, 2016a, 2016b). We extend this approach to ask related questions across two experiments to infer the parameters of the relevant prior knowledge. The major feature of this approach is that participants respond only to simple natural language questions rather than estimating numerical quantities for which complicated linking functions must be designed (Franke et al., 2016). The fully Bayesian language approach we pioneer here also provides a further test of the language model, which must predict data from two similar but distinct language experiments. Language is productive. We channel the productivity of language into experiment design, which can be harnessed to reduce uncertainty in the necessary parameters of language and cognitive models.

References

Bale, A. C. (2011). Scales and comparison classes. *Natural Language Semantics*, *19*, 169–190.

Barner, D., & Snedeker, J. (2008). Compositionality and Statistics in Adjective Acquisition: 4-year-olds Interpret Tall and Short Based on the Size Distributions of Novel Noun Referents. *Child Development*, *79*, 594–608.

Bergen, L., & Goodman, N. D. (2015). The strategic use of noise in pragmatic reasoning. *Topics in Cognitive Science*, *7*(2), 336–350.

Franke, M., Dablander, F., Scholler, A., Bennett, E., Degen, J., Tessler, M. H., ... Goodman, N. D. (2016). What does the crowd believe ? A hierarchical approach to estimating subjective beliefs from empirical data. In *Proceedings of the 38th annual meeting of the cognitive science society*.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*.

Goodman, N. D., & Stuhlmuller, A. (2014). The Design and Implementation of Probabilistic Programming Languages. `http://dippl.org`.

Lassiter, D., & Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positiveform adjectives. *Proceedings of SALT*, *23*(1), 587610.

Qing, C., & Franke, M. (2014a). Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. *Proceedings of SALT*, *24*.

Qing, C., & Franke, M. (2014b). Meaning and Use of Gradable Adjectives: Formal Modeling Meets Empirical Data. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*.

Ritchie, D., Stuhlmuller, A., & Goodman, N. D. (2016). C3: Lightweight incrementalized mcmc for probabilistic programs using continuations and callsite caching. In *AISTATS 2016*.

Rosch, E., & Mervis, C. B. (1975). Family Resemblances: Studies in the Internal Structure of Categories. *Cognitive Psychology*.

Schmidt, L. A., Goodman, N. D., Barner, D., & Tenenbaum, J. B. (2009). How Tall Is Tall? Compositionality, Statistics, and Gradable Adjectives. In *Proceedings of the 31st annual conference of the cognitive science society*.

Scholler, A., & Franke, M. (2015). Semantic values as latent parameters: Surprising *few* & *many*. In *Proceedings of SALT 45*.

Solt, S. (2009). Notes on the Comparison Class. In *International workshop on vagueness in communication*.

Solt, S., & Gotzner, N. (2012). Experimenting with degree. *Proceedings of SALT*, *22*, 166–187.

Tessler, M. H., & Goodman, N. D. (2016a). A pragmatic theory of generic language. *ArXiv*, *abs/1608.0*.

Tessler, M. H., & Goodman, N. D. (2016b). Communicating generalizations about events. In *Proceedings of the 38th annual meeting of the cognitive science society*.