Title

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

Permalink

https://escholarship.org/uc/item/0hq7w5bn

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 39(0)

Authors

Tessler, Michael Henry
Lopez-Brau, Michael
Goodman, Noah D.

Publication Date

2017

Peer reviewed

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]Dept. of Psychology, Stanford University, [2]Dept. 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 signal 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 formalize how world knowledge and listeners' internal models of speech production can drive the resolution of a comparison class in context. 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 with a follow-up 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). Four-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 in that context 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).

Any particular object of discourse, however, 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*. Why should one comparison class be preferred over another? To our knowledge, this question has not been addressed formally or empirically.[1] We propose that listeners actively combine category knowledge with pragmatic considerations to infer the comparison class implicitly 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 implicit comparison class.

We derive two novel **qualitative predictions** from this model. Saying "it's warm" in winter should signal it's warm *for winter* (as opposed to *for the year*) more so than saying "it's cold". The opposite relationship should hold in summer, where "it's cold" should signal it's cold *for summer* more so than "it's warm". This prediction is driven by the *a priori* probability that the adjective could apply to the class (e.g., the probability that a given day in winter is warm; Prediction 1). In addition, regardless of the season and the adjective form (e.g., "warm" or "cold"), listeners who expect speakers to be informative will prefer classes that are relatively specific (e.g., relative to *the current season* as opposed to *the whole year*), as they carry more information content (Prediction 2). We test these predictions by eliciting the comparison class using a paraphrase dependent measure (Expt. 1).

As with any Bayesian cognitive model, explicitly specifying relevant prior knowledge (e.g., beliefs about temperatures) is necessary for the model to make **quantitative predictions**. The current methodological standard is to measure beliefs by having participants estimate quantities or give likelihood judgments (Franke et al., 2016). We pursue a different methodology. The RSA model captures a productive fragment of natural language; thus, it makes predictions about a related natural language task (Expt. 2). Critically, we can use the model to predict natural language judgments that require the *same prior knowledge* as in Expt. 1 and use Bayesian data analysis to jointly infer the shared priors. This approach harnesses the productivity of language into experiment design and allows us to reconstruct priors without having participants engage in challenging numerical estimation tasks.

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 is probabilistically set with respect to a comparison class c via pragmatic reasoning (Lassiter & Goodman, 2013; see also Qing & Franke, 2014a):

[1] Theoretical work in semantics has instead focused on how information from a comparison class is used and what representations might be preferred (Bale, 2011; Solt, 2009).

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

$$S_1(u \mid x,\theta) \propto \exp\left(\alpha_1 \cdot \ln L_0(x \mid u,\theta)\right) \tag{2}$$

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

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 came from an approximately rational Bayesian speaker S_1 trying to inform a naive listener L_0, who in turn updates her prior beliefs $P_c(x)$ 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). θ comes from an uninformed prior and is resolved by the listener by reasoning about the likely states of the world $P_c(x)$ (e.g., possible temperatures) and the likelihood that a speaker would say the adjective given a state and a threshold $S(u \mid x,\theta)$. The prior distribution over world-states $P_c(x)$ 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., "...for the season"), we posit the listener infers the comparison class 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 how a rational speaker would behave in a given 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*). Crucially in this situation, the listener is aware that the target entity is a member of 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:

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

$$S_1(u \mid x,c,\theta) \propto \exp\left(\alpha_1 \cdot \ln L_0(x \mid u,c,\theta)\right) \tag{5}$$

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

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 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[2], 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.

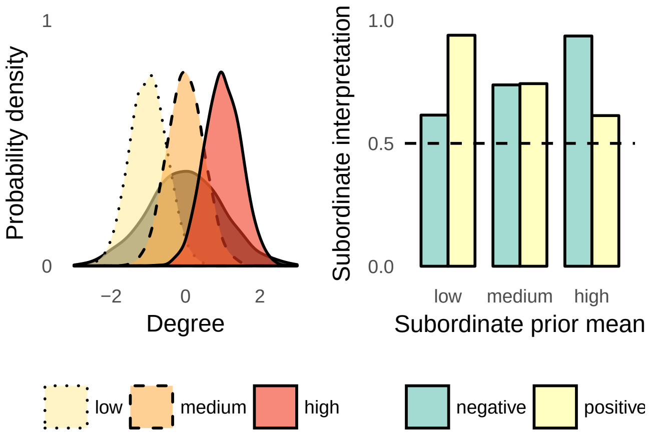

Figure 1: Left: Three hypothetical subordinate class prior distributions over a degree (fixing the superordinate class to be a unit-normal distribution, in grey). Right: Predicted listener inferences for an intended subordinate class interpretation given positive and negative form adjectives with different subordinate degree priors.

Qualitative model predictions Figure 1 (left) shows schematic superordinate and subordinate priors; e.g., temperatures over the whole year (super), in winter (low), fall (medium), and summer (high). The subordinate distributions have lower variance than the superordinate, and the "low" and "high" distributions have different means (e.g., temperatures in winter are expected to be lower and have lower variance than temperatures over the whole year).

Two intuitions explain the inferences of the pragmatic listener model (shown in Figure 1 right). First, certain classes are more or less likely to have an adjective felicitously apply. For example, any given day in winter is *less* likely to be warm than cold. Thus, hearing "it's warm" (a positive-form adjective) in winter (low prior) will signal it's warm *for winter* (the subordinate class) more so than hearing "it's cold" (negative-form), because it's more likely to be true (Prediction 1).

[2] Corpus accessed via `https://corpora.linguistik.uni-erlangen.de/cgi-bin/demos/Web1T5/Web1T5_freq.perl`. Due to potential polysemy and idiosyncracies of our experimental materials (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

Second, the amount of information conveyed by a vague utterance depends upon the variability in the comparison class. Comparison classes that have higher variance will result in relatively less information gain by the listener. All else being equal, listeners will prefer lower variance (e.g., subordinate) comparison classes because they are more informative (Prediction 2). Figure 1 (right) shows that subordinate class interpretations are above baseline regardless of the adjective polarity (positive or negative) or the mean of the subordinate prior (low, medium, high).

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 subordinate category. We test these two predictions in our first experiment.

Overview of data analytic approach As described above, specifying the relevant prior knowledge yields two free parameters per subordinate class. We will put priors over these parameters and infer their likely values using Bayesian data analysis. The data from the comparison class experiment (Expt. 1) would be insufficient, however, to reliably estimate all of the parameters of this data analytic model. To alleviate this, we use the same RSA model to predict additional data about related language use in the same domains (Expt. 2). Specifically, we gather judgments about adjectives when the comparison class is explicit: whether or not an adjective would apply to a subordinate member explicitly relative to the superordinate category (e.g., Is a day in winter warm relative to other days of the year?).

To model Expt. 2 data, we remove comparison class uncertainty by setting $P(c_{super}) = 1$, since the sentences provide an explicit comparison to the superordinate class. We 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 Expts. 1 and 2, and depend on the same background knowledge $P(x \mid c)$. We can thus use data from both experiments to jointly reconstruct the shared 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

Experiment 1 tests the qualitative predictions of the model. Experiment 2 collects further data about adjective usage in order to constrain the quantitative predictions of the RSA model, which will be used to predict data from both 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 left). This resulted in 30 unique items ({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*)

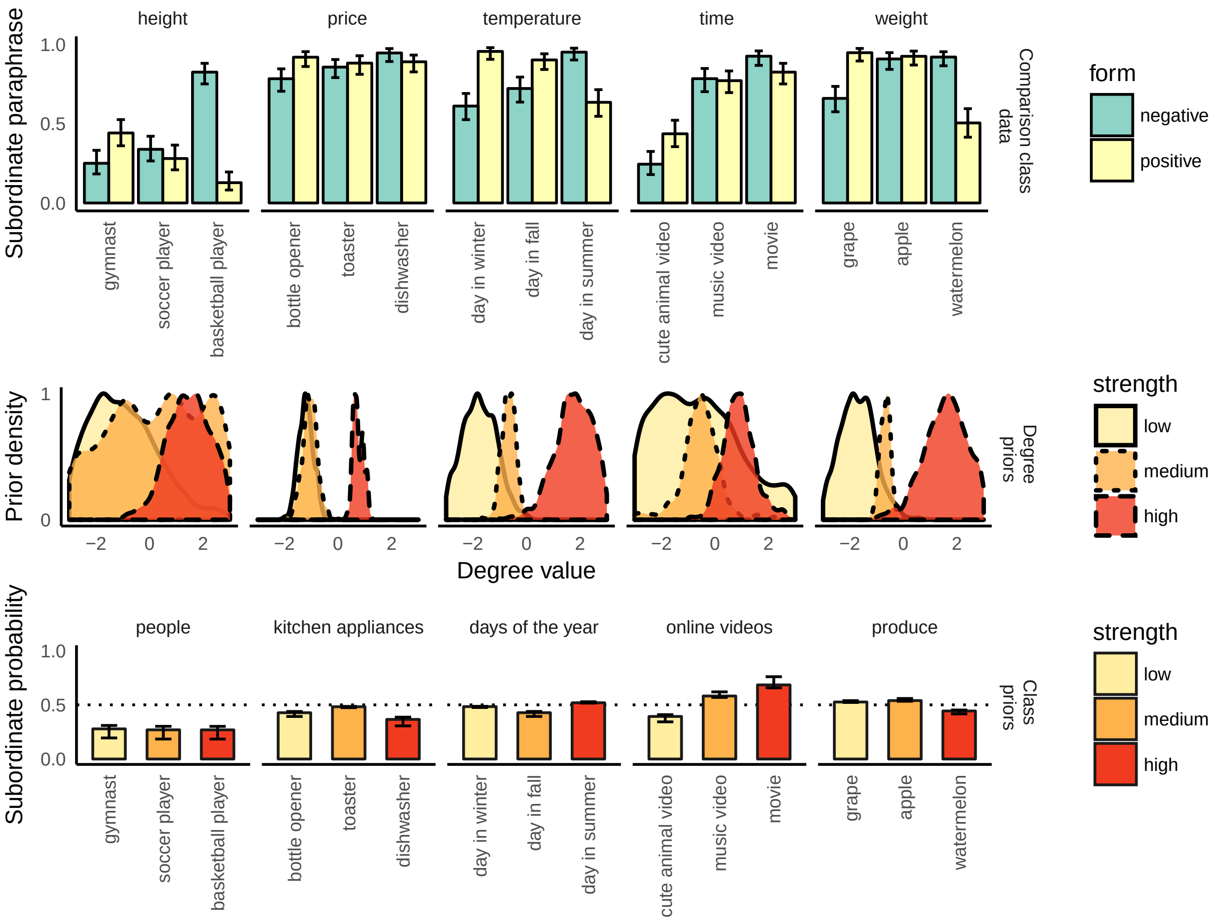

Figure 2: Empirical comparison class data, inferred world priors, and empirically derived comparison class priors. Top: Experiment 1 results. Comparison class judgments in terms of proportion judgments in favor of subordinate comparison class. Middle: Inferred prior distributions of world knowledge used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

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 salient.

Results We observed no systematic differences between participants' responses when the superordinate category was previously mentioned in the context and those when it was not; thus, we collapse across these two conditions for all analyses. Figure 2 (top) shows the proportion of participants choosing the *subordinate* paraphrase for each item, revealing considerable variability both *within*- and *across*- scales. The predicted effects are visually apparent within each scale (compare with Figure 1 right).

Our qualitative 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[3]. 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 main effects of form and strength were not significant.

We then test the simple effects. For items low on the degree scale (e.g., temperatures in winter), positive form adjectives were significantly more likely to imply subordinate 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; $\beta = -2.5$; $SE = 0.19$; $z = -13.15$). Participants reason pragmatically to resolve the comparison class, combining world knowledge with informativity as predicted by our model.

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

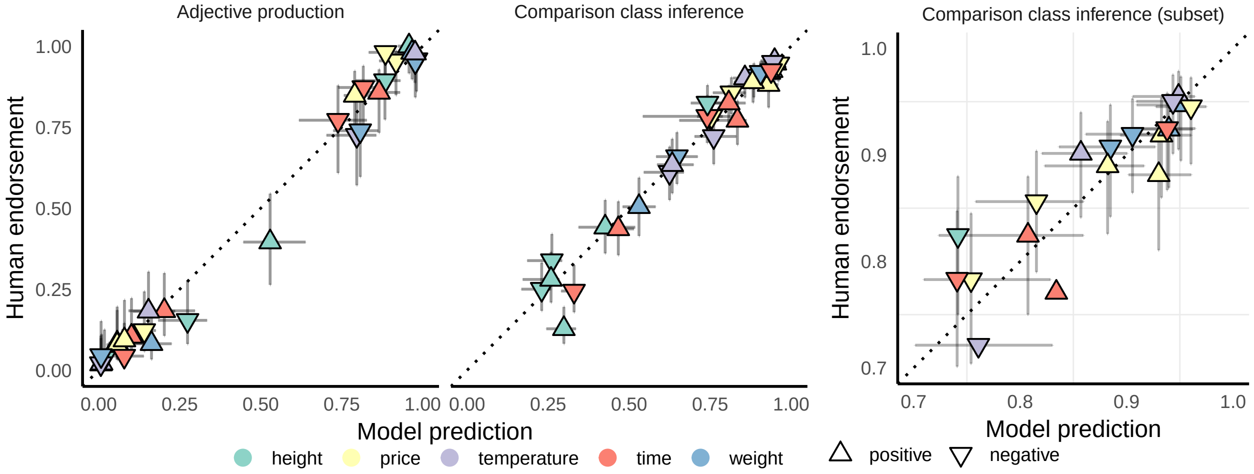

Figure 3: Human endorsement of subordinate comparison class paraphrases (middle; Expt. 1) and adjective sentences (left; Expt. 2) as a function of listener model L_1 and speaker model S_2 predictions, respectively. The right facet displays a subset of the paraphrase data (Expt. 1) to reveal good quantitative fit even in a small dynamic range. Error bars correspond to 95% Bayesian credible intervals.

Experiment 2: Adjective endorsement

In this experiment, we collected data about adjective endorsement that would require the same prior knowledge relevant for Expt. 1. We use this data to further constrain the RSA model's quantitative predictions.

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 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 judgments in this experiment were consistent with the *a priori* ordering of the subordinate categories on the degree scale. On the y-axis of Figure 3 (left), 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 RSA listener (Eq. 4) and speaker (Eq. 7) models make quantitative predictions about comparison class interpretation and adjective endorsement, respectively. We construct a single data-analytic model with each of these RSA components as sub-models in order to make quantitative predictions about the data from both of our experiments.

The listener and speaker sub-models share their prior world knowledge $P(x \mid c)$ (e.g., temperatures in winter), described in the **Degree Priors** section. We put the same priors over the parameters of each subordinate distribution: $\mu \sim \text{Uniform}(-3,3)$, $\sigma \sim \text{Uniform}(0,5)$, since they have standardized units. The comparison class prior $P(c)$ in Eq. 4 scales the empirical frequency $\hat{f}$ by a free parameter, which we give the following prior: $\beta \sim \text{Uniform}(0,3)$.

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. The pragmatic listener L_1 model (Eq. 4) has one speaker optimality: α_1^1. The pragmatic speaker S_2 model (Eq. 7) has two speaker optimality parameters: $\{\alpha_1^2, \alpha_2^2\}$. We use priors consistent with the previous literature: $\alpha_1 \sim \text{Uniform}(0,20)$, $\alpha_2 \sim \text{Uniform}(0,5)$

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, we collecting 2 chains of 50k iterations (after 25k burn-in) using an incrementalized version of MCMC (Ritchie, Stuhlmuller, & Goodman, 2016).

Results The full model's posterior over the RSA and data-analytic parameters were consistent with 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 $\alpha_1^1 = 1.6[1.1,2.5]$, $\beta = 0.13[0.11,0.19]$. Model parameters specific to the S_2 model used for Expt. 2: $\alpha_1^2 = 3.5[0.6,13.2]$, $\alpha_2^2 = 3.2[2.6,3.8]$. The inferred distributions corresponding to subordinate class priors were consistent with the *a priori* ordering of these subordinate classes (low, medium, high) used in these tasks (Figure 2 middle).

Finally, the full model's posterior predictive distribution does an excellent job at capturing the quantitative 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 above 0.5. Even for these fine-grained differences, the model does a good job at explaining the quantitative variability in participants' data (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 a speaker might be implicitly referring to or comparing against. Here, we investigate the flexibility in the class against which an entity can be implicitly compared.

We introduced a minimal extension to an adjective interpretation Rational Speech Act model to allow it to flexibly reason about the comparison class. This model made two novel predictions about how listeners should prioritize one class over another. It also made quantitative predictions about how background knowledge about the degree scale should inform this inference in a graded fashion. Both qualitative predictions of the model were borne out in our first experiment, and the quantitative predictions were confirmed using a novel data analytic technique. To our knowledge, this is the first experiment to demonstrate how reference classes for adjective interpretation can adjust based on world knowledge.

We observe in our modeling results for Expt. 1 that a uniform prior distribution over the experimentally supplied comparison class alternatives is unlikely (Figure 2 bottom). For example, the comparison class of "people" for heights of individuals is relatively more salient than the class of "produce" for the weights of fruits and vegetables. We used the frequency of the class in a corpus as a proxy for their prior probability $P(c)$, which was sufficient to account for differences in baseline class probability both *between*- and *within*-scales.

Corpus frequency is a composite measurement of factors relevant for speech production. Its utility in this model suggests that utterances without an explicit comparison class (e.g., "It's warm outside") may in fact be incomplete sentences, in a way analogous to sentence fragments studied in noisy-channel models of production and comprehension (Bergen & Goodman, 2015). Another (non-mutually exclusive) possibility is that the comparison class prior reflects basic-level effects in categorization (Rosch & Mervis, 1975). Future work should attempt to understand 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 related language experiments. In previous work, we have attempted to measure prior knowledge by decomposing what would be a single, implicitly multi-layered, numerical estimation question into multiple simpler questions. Then, we construct a Bayesian data analytic model to back out the prior knowledge (Tessler & Goodman, 2016a, 2016b). We extend this approach by using the same core RSA model to model behavior across two language experiments. The major feature of this method is that participants respond only to simple, natural language questions rather than estimating numerical quantities for which complicated linking functions must be designed (e.g., Franke et al., 2016). The fully Bayesian language approach we pioneer here also provides a further constraint on the language model, which must predict data from two similar but distinct language experiments. The productivity of natural language can thus be harnessed to productively design experiments that further constrain and test computational models of language and cognition.

Acknowledgements

The authors would like to thank Ali Horowitz for help in stimuli design. This work was supported in part by NSF Graduate Research Fellowship DGE-114747 to MHT, a Stanford CSLI Summer Internship for MLB, and a Sloan Research Fellowship, ONR grant N00014-13-1-0788, and DARPA grant FA8750-14-2-0009 to NDG.

References

Bale, A. C. (2011). Scales and comparison classes. *Natural Language Semantics*, *19*(2), 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*(3), 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*, *20*(11), 818–829.

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 positive-form adjectives. In *Semantics and linguistic theory* (Vol. 23, pp. 587–610).

Qing, C., & Franke, M. (2014a). Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. In *Semantics and linguistic theory* (Vol. 24, pp. 23–41).

Qing, C., & Franke, M. (2014b). Meaning and Use of Gradable Adjectives: Formal Modeling Meets Empirical Data. In *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*, *7*(4), 573–605.

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 *Semantics and linguistic theory* (Vol. 25, pp. 143–162).

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

Solt, S., & Gotzner, N. (2012). Experimenting with degree. In *Semantics and linguistic theory* (Vol. 22, pp. 166–187).

Tessler, M. H., & Goodman, N. D. (2016a). A pragmatic theory of generic language. *ArXiv Preprint ArXiv:1608.02926*.

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