The importance of alternatives in scalar implicature

Benjamin Peloquin
bpeloqui@stanford.edu
Department of Psychology
Stanford University

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Abstract

Succesful communication regularly requires listeners to make pragmatic inferences - enrichments beyond the literal meaning of a speaker's utterance. For example, listeners routinely enrich the meaning of "some" to "some, but not all" when interpreting a sentence such as "Bob ate some of the cookies." A Gricean account of this phenomena assumes the presence of *salient alternatives* with varying degrees of informativity. "Some," in the example above, is enriched to "some, but not all" in the presence of the stronger alternative "all." Empirical evidence for the presence of such alternatives and accounts of their effect on implicature has been limited. Our current work explores the role different scale representations may play in scalar implicature using emperical measures of literal semantics and a Bayesian model of implicature generation. Comparisons with human judgments indicate that pragmatic inference may rely on fairly complete alternative sets rather than the typical entialment only scales assumed in most research to date.

Keywords: pragmatics; scalar implicature; bayesian modeling

Introduction

Humpty Dumpty Quote...

Successful communication regularly requires listeners to make inferences that go beyond the literal semantic content of a speaker's utterance. "Scalar implicature" is a hallmark of such pragmatic enrichment. For example, listeners routinely enrich the meaning of the scalar item "some" to "some, but not all" in sentences like "Bob ate some of the cookies" (van Tiel, 2014). A Gricean account of this phenomena assumes listeners reason about the intended meaning of a speaker while incorporating knowledge about a) alternative scalar items a speaker could have used (such as "all") and b) the relative informativity of using such alternatives (Grice, 1975). Under the Gricean account, a listener will infer that the speaker must have intended that Bob did not eat "all" the cookies because it would have been *underinformative* for the speaker to use "some" when "all" could have been used.

The presence of "salient" alternatives is critical to the standard Gricean account of scalar implicature (Grice, 1975). Recent experimental evidence appears to support this. Using a gumball paradigm, Degen & Tanenhaus (2014) demonstrated that the scalar item *some* was judged less appropriate (natural) for small sets of items when exact numbers were seen as viable alternatives. This finding points to the impact a particular alternative set can have on judgments regarding scalar appropriateness. Using a fundamentally differnt approach, Franke (2014) formalized the Gricean account of implicature using a Bayesian model, modeling graded typicality judgments and scalar implicature. Results supported 'none', 'one' and 'all' as the most serious (salient) alternatives to "some."

Taken together, each of the previous studies signal the importance of scale representation (set of available alternatives) in scalar implicature. However, like many recent investigations which have focused only on a small subset of scalar items, both Degen & Tanehaus (2014) and Franke (2014) focus exclusively on the scalar family "some/all." This lack "Scalar Diversity" (van Tiel, 2014) makes generalization across different scalar families problematic. van Tiel (2014), found substantial variation in the rate of upper-bounded construals among over 40 different scalar pairs. This finding suggests that implicature may differ widely between different scales. We adopt a "Scalar Diversity" approach in our experimental design in order to extend our investigation beyond "some/all" scalar items. In the following set of studies we examine implicature for "some/all" and five additional scalar families from a range of grammatical classes (see Figure 5 for a complete set of scalar items used). The set of scalar items for this study were chosen from among the 40 used in van Tiel (2014).

Consider a scalar item other than "some" - in the context of a scalar item such as "good", the idea of set of lexical alternatives certainly xfeels intuitive ("bad", "excellent", etc.). But measuring the extent to which these alternatives are salient is not so clear. Nor is how to measure their impact on pragmatic intepretations of the scalar item "good." In an experimental setting, querying a participant about the saliency of a particular alternative is problematic because the alternative must be made salient during the query. In the spirit of Franke (2014) and Degen & Tanenhaus (2014), we pair a computational model of implicature with empirical measurments of linguistic material. This combination allows us to simulate the effect of various scale representations on implicature. In particular, we adopt a Bayesian model of implicature and a food-review paradigm to empirically quantify literal semantics and pragmatic judgments for our scalar stimuli.

Bayesian models of pragmatic enrichment have accurately predicted human judgments in ad-hoc (Frank & Goodman, 2012) and embedded (Stulhumuller & Goodman, 2014) implicature settings. Our current model follows both these studies and is based on Rational Speech-act theory (RSA). RSA frames language understanding as a special case of social cognition (Stulhumuller & Goodman, 2014), in which listeners and speakers reason about one-another. In this framework, a listener uses Bayesiain inference to interpret the intended meaning m of a speaker who has made an utterance u. We will make the model specifications explicit later in the paper.

In the following study we investigate the role of scale representation on scalar implicature. To do so we combine a computational framework with empirical measurements to simulate the type of scale representations that might be avail-

able to a listener. In the following sections we outline the specifics of the model and the particular components for which we need to to make empirical measurements. We then outline a series of experiments meant to a) quantify literal semantics for otherwise ambiguous scalar items, b) measure pragmatic judgments and c) generate a set of plausible alternatives for our set of scalar pairs taken from van Tiel (2014). In particular, we adopt a food review paradigm throughout our empirical data gathering, quantifying linguistic meaning (both literal semantics and pragmatic judgments) as distributions over star ratings. Finally, we simulate the impact of three different scale representations, "entailment", "mid" and "symmetric" on implicature generation, comparing model predictions with pragmatic judgments captured empirically. Results indicate that humans may employ more "fully specified" scale representatinos, relying on the presence of multiple alternatives during pragmatic enrichment.

Modeling implicature: Rational Speech-act theory

We adopt a Bayesian model of scalar implicature known as Rational Speech-act theory (RSA). In particular, RSA frames language understanding as a special case of social cognition (Frank & Goodman, 2012; Stuhmuller & Goodman, 2014) in which listener and speaker agents reason about one another. We focus on the problem of a listener inferring the meaning of a speaker's utterance. Let u be an observed utterance made by our speaker with an intended meaning m. We assume that the listener has access to a space of possible meanings M and some model relating meanings $m \in M$ to u.

Upon hearing an utterance u the Listener agent evaluates all candidate word meanings, computing their posterior probabilities $p_{L1}(m|u)$. This quantity is proportional to the product of the prior probability $p(m)$ of a particular meaning and the likelihood $p(u|m)$.

$$p_{L1}(m|u) = \frac{p(u|m)p(m)}{p(u)}$$
$$= \frac{p(u|m)p(m)}{\sum_{m \in M} p(u|m)}$$

The **prior** $\mathbf{p(m)}$ represents the listener's expectations about plausible meanings (star ratings), *independent* of the utterance u. To avoid biasing the data we assume priors over meanings are uniform, however future investigations may want to experiment with different priors, either by inferring them or using empirical measures

$\mathbf{p(u|m)}$ quantifies the **likelihood** that a speaker would make a particular utterance u given an intended meaning m. Our model assumes that speakers choose words to maximize the utility of an utterance in context. Utility is operationalized as the informativity of a particular utterance (suprisal) minus a cost. In particular, the speaker chooses an utterance in order to maximize the utility of a listener who interprets her utterance literally $p_{L0}(m|u)$. L_0 denotes a literal listener interpretation.

$$p(u|m) = \frac{e^{-\alpha(-log(p_{L0}(m|u))-cost(u))}}{\sum_{m \in M} e^{-\alpha(-log(p_{L0}(m|u))-cost(u))}}$$

The **posterior** $\mathbf{p_{L1}(m|u)}$ quantifies a listener's pragmatic enrichment of an utterance u as degree of belief that m is the intended meaning of the speaker, given an utterance u. L_1 denotes a pragmatic listener interpration.

Populating the model: Measuring literal semantics and pragmatic judgments

In order to simulate the role of scale representation on implicature we pair a computational framework with empirical data. In particular we use experimental tools to populate three components of the model. First, to measure otherwise ambigous literal semantics $\mathbf{p_{L0}(m|u)}$ (our "Scalar Diversity"" approach). Second, to generate a set of plausible alternatives to the original scalar pairs taken from van Tiel (2014). Lastly, to obtain human pragmatic judgments for model comparison $\mathbf{p_{L1}(m|u)}$.

Starting with the first component, consider the scalar items "some" and "all." Both are convenient for modeling in a computational framework like the one we've proposed above. Their literal semantics $\mathbf{p_{L0}}$ are easily quantified over sets.

some of the A's are B's.

$$[[some]] = \{< A,B >: |A \cap B| \geq 1\}$$

all of the A's are B's.

$$[[all]] = \{< A,B >: \frac{|A \cap B|}{|A|} == 1\}$$

The same cannot be said for items such as "good / excellent", "memorable / unforgettable" or "liked / loved."

In the following set of studies we adopt a food-review paradigm to measure literal semantics and pragmatic judgments empirically. In order to capture graded typicality judgments as in Degen & Tanenhaus (2014) we use a compatiblity measure dependent variable, pairing scalar items with star-ratings in our literal listener task (Experiments 1a, 3a and 4). Data from these experiments is used to approximate the literal listener semantics for a given scalar item u as a distribution over star-ratings $m \in \{1 - 5 \text{ stars}\}$. Similarly, pragmatic judgment data from Experiments 1b and 3b serve as our primary points of comparison with model predictions $\mathbf{p_{L1}(m|u)}$, again quantifying judgments as distributions over star-ratings. In the following section we take the read through our experiments.

Experiment 1a,b: Entailment scales

Experiment 1a and 1b were conducted to approximate literal listener semantic distributions (1a) and pragmatic judgments (1b) for five pairs of scalar items taken from van Tiel (2014)

Figure 1: The left panel shows a trial from Experiment 1a with the target scalar 'liked'. Participants responses to the binary dependent variable are used to quantify literal semantics. The right panel shows a trial from Experiment 1b with the target scalar 'liked'. Participants were asked to generate the star rating they think the speaker likely intended, given the target scalar (in this case 'liked').

(see figure 5 for stimuli). Each pair consists of a "weaker" scalar term (e.g. "some") and a stronger scalar term (e.g. "all"). We are calling these "entialment scales"", because they encode the assumption that a listener need only have a the stronger alternative present to generate an enriched interpretation of the weaker term.

Methods

Participants Participants for both tasks were recruited on Amazon Mechanical Turk and paid $0.20 for their participation. Thirty participants were recruited for Experiment 1a. Data for N participants was thrown out after participants either failed to pass two training trials or were not native English speakers, leaving a total sample of N. Fifty participants were recruited for experiment 1b. Data for N participants was thrown out after participants either failed to pass two training trials or were not native English speakers, leaving a total sample of N.

Design and procedure The left panel of Figure 1 shows a screen shot of a trial from Experiment 1a. Participants were presented with a target scalar item and a star-rating (between 1-5 stars) and asked to judge the compatibility of the scalar item and star-rating. Compatibility was assessed through a binary "yes/no" response to a question of the form "Do you think that the person ___ed the food?" where a target scalar was presented in the "___ed." Each participant was presented all scalar item and star-rating combinations with randomization.

The right panel of Figure 1 shows a screen shot of a trial from Experiment 1b. On each trial, participants were presented with a one-sentence prompt containing a target scalar item such as "Someone said they ___ed the food." They were then asked to generate a star-rating according to what they think the reviewer likely gave. Each participant was presented all scalar items with randomization.

Results and Discussion

The left panel of figure two shows literal semantics for each of the scalar terms "liked" and "loved" as distributions over star-ratings.

Chi-squared test? Null results, not really important here anyway...

Experiment 2: What are the alternatives?

In Experiment 2 we asked participants to generate plausible alternatives for the scalar items presented in Experiment 1.

Methods

Participants Using Amazon's Mechanical Turk, N workers were paid $0.20 to participate. n workers were dropped from the analysis because they either failed to pass training trials or were not native English speakers. The final sample was N workers, all of whom were naive to the purpose of the experiment.

Design and procedure Participants were presented a target scalar term embedded in a sentence such as, "In a recent restaurant review someone said they thought they —-ed the food." Participants were then asked to generate plausible alternatives by responding to the question, "If they'd felt differently about the food, what other words could they have used instead of —-?" and asked to generate three alternatives.

Results and Discussion

Figure 3 plots the combined counts of alternatives generated for the scalar items "liked" and "loved." Alternative distributions for the other scalar pairs (e.g.. "good/excellent", "memorable/unforgettable") were similarly long-tailed. In Experiments 3a,b we took the top two alternatives generated for each scalar item to enrich the entailment only scales from Experiments 1a,b. In experiment 4 we further enriched the scales from Experiments 3a,b with a neutral valence scalar chosen from the alternative sets.

Experiment 3a,b: Incorporating top alternatives

Experiment 3a,b are identical to Experiments 1a,b except the set of target scalar was expanded to include the top two alternatives generated for each scalar family. The orange colored items in table 1 denote additional scalar items added in Experiments 3a,b.

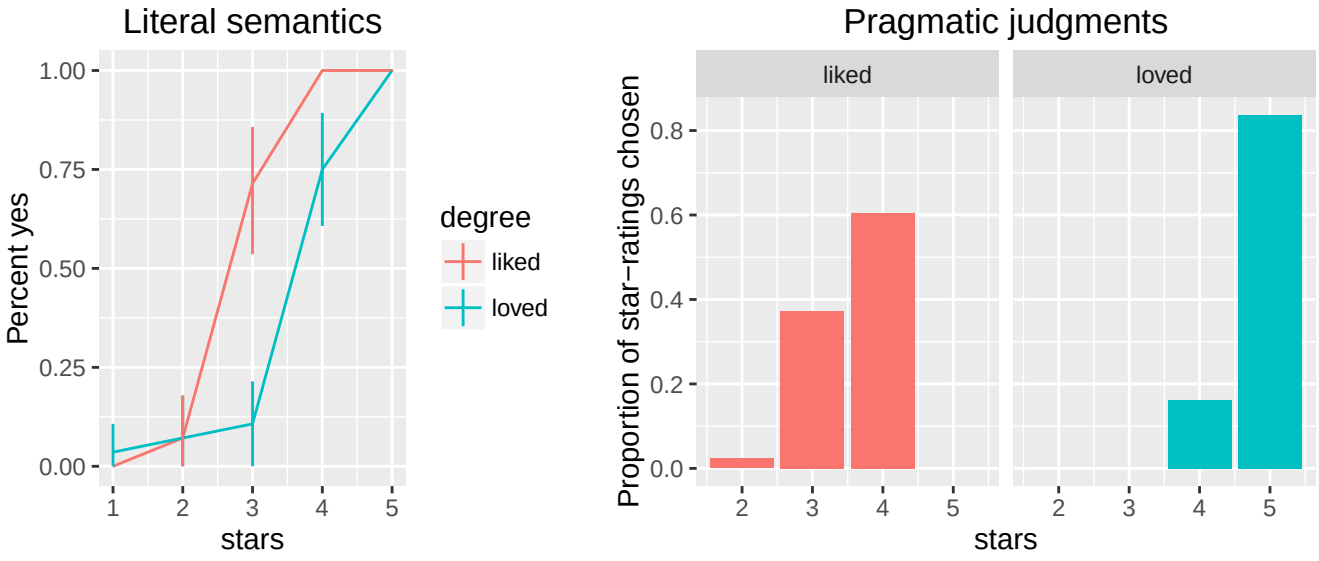

Figure 2: The left panel plots literal semantic distributions for the scalar pair 'liked/loved'. The y-axis is the percentage of 'yes' responses for a scalar term that is compatibible with the number of stars on the x-axis (ie 100% of respondents beleived that 'liked' was compatible with both 4 and 5 stars). Erro bars are 95% confidence intervals. The right panel plots pragmatic judgments for the scalar pair 'liked/loved'. The y-axis is the proportion of selection of the star-rating on the x-axis (ie over 80% of judgments when prompted with 'loved' were 5-stars).

Participants

Participants for both studies were recruited on Amazon Mechanical Turk and paid $0.20 for their participation. Thirty participants were recruited for Experiment 3a (Literal Listener task). Data for N participants was thrown out after participants either failed to pass two training trials or were not native English speakers, leaving a total sample of N.

Fifty participants were recruited for Experiment 3b (Pragmatic Listener task). Data for N participants was thrown out after participants either failed to pass two training trials or were not native English speakers, leaving a total sample of N.

Procedure

The procedure of Experiments 3a,b follow the same form as Experiments 1a,b with the expanded set of target scalar items.

Results and Discussion

Distributional differences.

Experiment 4: Symmetric scales

Experiment 4a is identical to Experiments 1a and 3a except the set of target scalar was expanded to include a neutral valence scalar item for each scalar family. The purple colored items in Table 1 denote additional scalar items added in Experiments 4.

Participants

Participants for both studies were recruited on Amazon Mechanical Turk and paid $0.20 for their participation. Thirty participants were recruited for Experiment 4 (Literal Listener task). Data for N participants was thrown out after participants either failed to pass two training trials or were not native English speakers, leaving a total sample of N.

Procedure

The procedure of Experiment 4 follow the same form as Experiments 1a and 3a with the addition of a neutrally valence scalar item.

Results and Discussion

Distributional differences.

Temp section Model was here

Results Table here

Figure 4 plots model / human correlation for our three simulations. Alpha was tuned separately for each of these runs...

General Discussion

By varying the type of scale representations available to our Bayesian model we investigated the effects of alternatives on scalar implicature. Model fit with human judgement was significantly improved by the inclusion of alternatives beyond the typical "strong" and "weak" scalar items. In fact, we found that both neutral and negative valence scalar items contribute to human-like implicature generation within our framework.

\begin{CodeChunk} \begin{figure}[h]

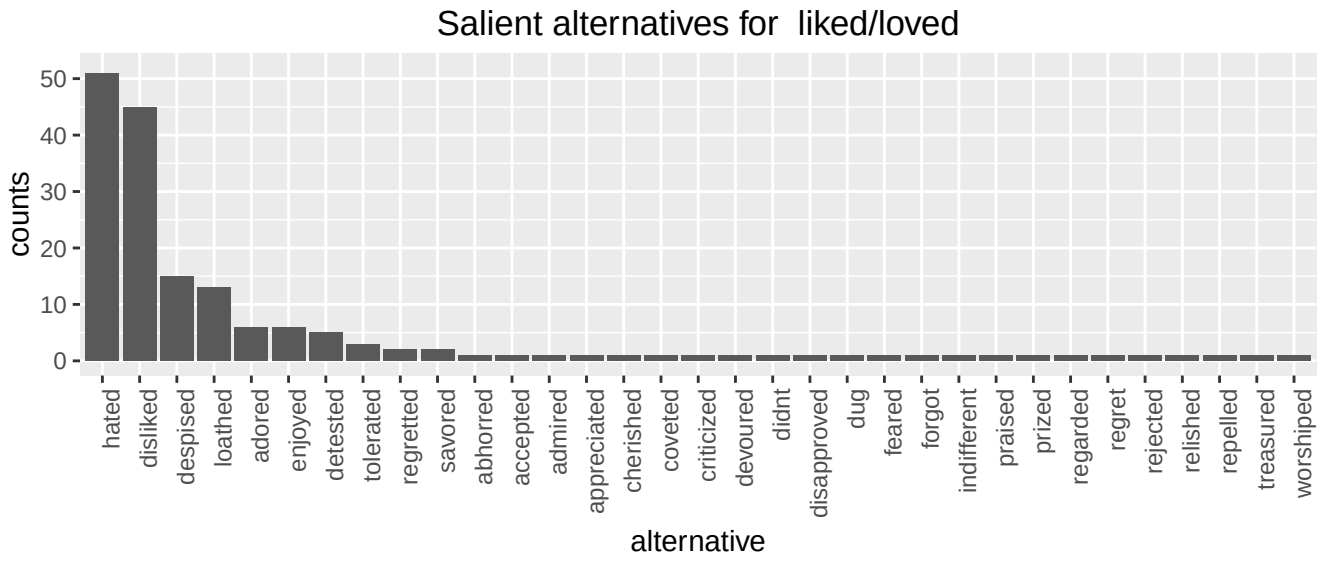

Figure 3: Caption goes here

Alternative sets				
good / excellent	liked / loved	memorable / unforgettable	palatable / delicious	some / all
excellent	loved	unforgettable	delicious	all
good	liked	memorable	palatable	most
okay	felt indifferent about	ordinary	mediocre	some
bad	disliked	bland	gross	little
horrible	hated	foregettable	disgusting	none

Entailment items used in Experiments 1a,b
Top two alternatives added in Experiments 3a,b
Neutral item added in Experiment 4

{

}

\caption[This table shows the stimuli used in Experiments 1a,b, 3a,b and 4]{This table shows the stimuli used in Experiments 1a,b, 3a,b and 4. Colors denote the additions made in each experiment. For example, in Experiment 1a we measured literal semantics for 'liked/loved'. In Experiment 3a,b we extended this group to 'textit{hated/disliked/liked/loved'. In Experiment 4 we extended this grou to 'hated/disliked/felt indifferent about/liked/loved'.} \end{figure} \end{CodeChunk}

Acknowledgements

Thanks to NSF BCS XYZ. Thanks to Michael Franke, Judith Degen, and Noah Goodman.

References

use a food review paradigm in which 'meaning' for a given scalar item was quantified as a distribution over star-ratings. Quantifying both literal semantic content and pragmatic judgments as distributions over stars provides an important interface to our model, especially for scalar items with ambigous literal semantics (e.g. items such as "liked/loved" or "memorable/unforgettable"). In addition to quantifying literal semantics and pragmatic judgments in Experiments 1a,b and Experiments 3a,b, Experiment 2 used a modified cloze task, based on van Tiel (2013, Experiment 2???), to generate a set of plausible alternatives to the scalar items used in Experiments 1a,b. We extend the set of alternatives measured in Experiments 3a,b and 4 using data from Experimeng 2. Additionally, by looking at the relative frequencies of alternatives generated in Experiment 2 we were able to compute a saliency measure for each alternative. Using the literal semantic data from Experiments 1a, 3a and 4, we modeled pragmatic enrichment using RSA and compared model predictions to human judgments obtained in Experiments 1b and 3b. Model performance was significantly improved with the incorporation of alternatives. This result suggests that more "fully specified" scale representations may be active for human participants while making pragmatic judgments for scalar items.

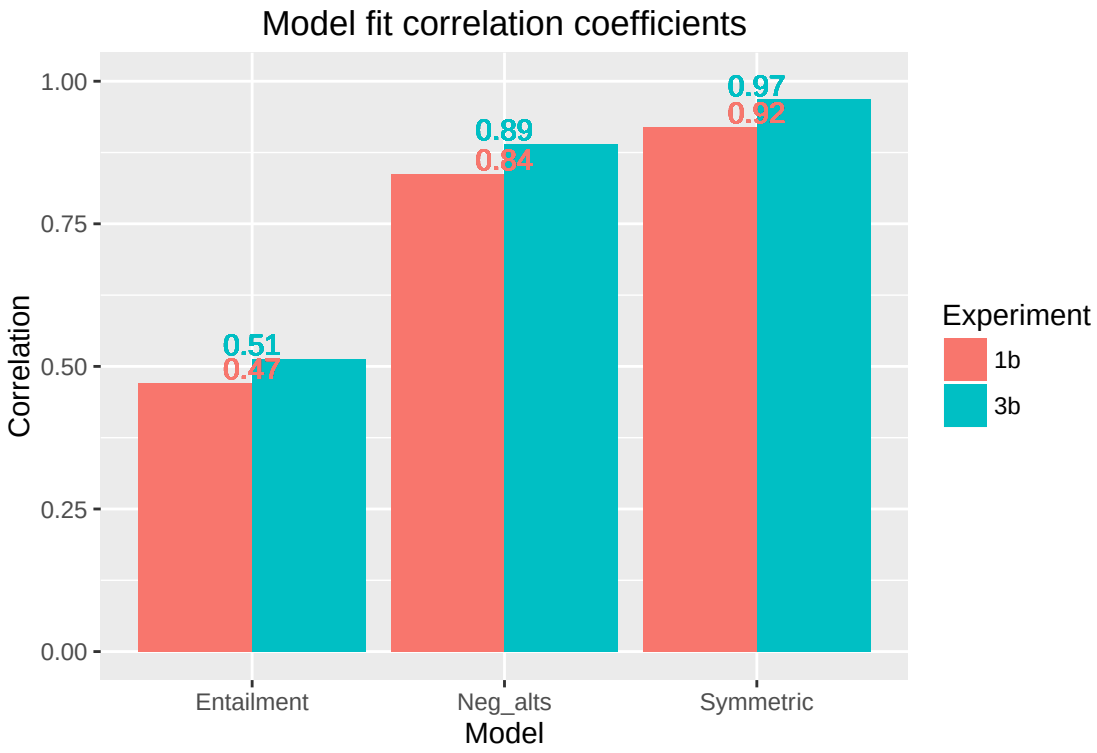

Figure 4: The left panel shows improved model fit as scale representations are enriched with more scalar items. Correlations are coputed using pragmatic judgment data from Experiments 1b and 3b. The right panel plots model predictions using full symmetric scales versus human judgments from Experiment 3b.