Determining the alternatives for scalar implicature

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

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

Abstract

Successful communication regularly requires listeners to make pragmatic inferences - enrichments beyond the literal meaning of a speaker's utterance. For example, when interpreting a sentence such as "Alice ate some of the cookies," listeners routinely infer that Alice did not eat all of them. A Gricean account of this phenomena assumes the presence of alternatives (like "all of the cookies") with varying degrees of informativity, but it remains an open question precisely what these alternatives are. We use a computational model of pragmatic inference to test hypotheses about how well different sets of alternatives allow us to predict scalar implicature performance across a range of different scales. Our findings suggest that human comprehenders likely consider a much broader set of alternatives beyond those entailed by the initial description.

Keywords: pragmatics; scalar implicature; bayesian modeling

Introduction

How much of what we mean comes from the words that go unsaid? As listeners, our ability to make precise inferences about a speaker's intended meaning in context is an indispensable component of successful communication. For example, listeners commonly enrich the meaning of the scalar item *some* to *some but not all* in sentences like "Alice ate some of the cookies" (Grice, 1975; Horn, 1984; Levinson, 2000). These inferences, called *scalar implicatures*, have been an important test case for understanding pragmatic inferences more generally. A Gricean account of this phenomena assumes listeners reason about the meaning the speaker intended by 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). According to this account, a listener will infer that the speaker must have intended that Alice did not eat *all* the cookies because it would have been underinformative for the speaker to use *some* when *all* could have been used.

But what are the alternatives that should be considered in this computation more generally? Under classic accounts of implicature, listeners consider only those words whose meaning would entail the word that is actually sent (Horn, 1972), and these alternatives enter into conventionalized or semi-conventionalized scales (Levinson, 2000). For example, because *all* entails *some*, and hence is a "stronger" meaning, *all* should be considered as an alternative to *some* in implicatures. Similar scales exist for non-quantifier scales, e.g. *loved* entails *liked* (and hence "I liked the movie" implicates that I didn't love it).

Recent empirical evidence has called into question whether entailment scales are all that is necessary for understanding scalar implicature. For example, Degen & Tanenhaus (2015) demonstrated that the scalar item *some* was judged less appropriate when exact numbers were seen as viable alternatives. And in a different paradigm, Tiel (2014) found converging evidence that *some* was judged to be atypical for small quantities. These data provide indirect evidence about a broader set of alternatives: since *some* is logically true of sets with one or two members, these authors argued that the presence of salient alternatives (the words *one* and *two*) reduced the felicity of *some* via a pragmatic inference.

By formalizing pragmatic reasoning, computational models can help provide more direct evidence about the role that alternatives play. The "rational speech act" model (RSA) is one recent framework for understanding inferences about meaning in context (Frank & Goodman, 2012; N. D. Goodman & Stuhlmller, 2013). RSA models frame language understanding as a special case of social cognition, in which listeners and speakers reason recursively about one another's goals. In the case of scalar implicature, a listener makes a probabilistic inference about what the speaker's most likely communicative goal was, given that she picked the quantifier *some* rather than the stronger quantifier *all*. In turn, the speaker reasons about what message would best convey her intended meaning to the listener, given that he is reasoning in this way. This recursion is grounded in a "literal" listener who reasons only according to the basic truth-functional semantics of the language.

Franke (2014) used an RSA-style model to assess what alternatives a speaker would need to consider in order to produce the typicality/felicity ratings reported by Degen & Tanenhaus (2015) and Tiel (2014). In order to do this, Franke (2014)'s model assigned weights to a set of alternative numerical expressions. Surprisingly, along with weighting *one* highly (a conclusion that was supported by the empirical work), the best-fitting model assigned substantial weight to *none* as an alternative. This finding was especially surprising considering the emphasis of standard theories on scalar items that stand in entailment relationships with one another (e.g. *one* entails *some* even if it is not classically considered to be part of the scale).

In our current work, we pick up where these previous studies left off, considering the set of alternatives for implicature using the RSA model. To gain empirical traction on this issue, however, we broaden the set of scales we consider. Our inspiration for this move comes from work by Van Tiel, Van Miltenburg, Zevakhina, & Geurts (2014), who examined a phenomenon that they dubbed "scalar diversity," namely the substantial difference in the strength of scalar implicature across a variety of scalar pairs (e.g. *liked/loved,* or *palat-*

able/delicious.). Making use of some of this diversity allows us to investigate the ways that different alternative sets give rise to implicatures of different strengths across scales.

We begin by presenting the computational framework we use throughout the paper. We next describe a series of experiments designed to measure both the literal semantics of a set of scalar items and comprehenders' pragmatic judgments for these same items. These experiments allow us to compare the effects of different alternative sets on our ability to model listeners' pragmatic judgments. To preview our results: we find that standard entailment alternatives do not allow us to fit participants' judgments, but that expanding the range of alternatives empirically (by asking participants to generate alternative messages) allows us to model listener judgments with high accuracy.

Modeling Implicature Using RSA

We begin by giving a brief presentation of the basic RSA model. This model simulates the judgments of a pragmatic listener who wants to infer a speaker's intended meaning m from her utterance u. For simplicity, we present a version of this model in which there is only full recursion: that is, the pragmatic listener reasons about a pragmatic speaker, who in turn reasons about a "literal listener." We assume throughout that this computation takes place in a signaling game (Lewis, 1969) with a fixed set of possible meanings $m \in M$ and a fixed possible set of utterances $u \in U$, with both known to both participants. Our goal in this study is to determine what utterances fall in U.

In the standard RSA model, the pragmatic listener (denoted L_1), makes a Bayesian inference:

$$p_{L1}(m \mid u) \propto p_{S_1}(u \mid m)p(m)$$

In other words, the probability of a particular meaning given an utterance is proportional to the speaker's probability of using that particular utterance to express that meaning, weighted by a prior over meanings. This prior represents the listener's *a priori* expectations about plausible meanings, independent of the utterance. Because our experiments take place in a context in which listeners should have very little expectation about which meanings speakers want to convey, for simplicity we assume a uniform prior $p(m) \propto 1$.

The pragmatic speaker in turn considers the probability that a literal listener would interpret her utterance correctly:

$$p_{S_1}(u \mid m) \propto p_{L_0}(m \mid u)$$

where L_0 refers to a listener who only considers the truth-functional semantics of the utterance (that is, which meanings the utterance can refer to).

This model of the pragmatic speaker (denoted S_1) is consistent with a speaker who chooses words to maximize the utility of an utterance in context (Frank & Goodman, 2012), where utility is operationalized as the informativity of a particular utterance (surprisal) minus a cost:

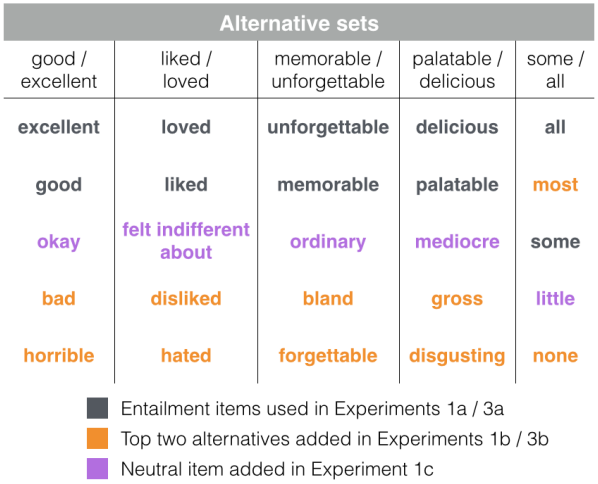

Figure 1: Stimuli for Experiments 1, 2, and 3

$$p_{S_1}(u \mid m) \propto e^{-\alpha(-log(p_{L_0}(m|u))-C(u))}$$

where $C(u)$ is the cost of a particular utterance, $-log(p_{L_0})$ represents the *surprisal* for the literal listener (the information content of the utterance), and α is a parameter in a standard choice rule. If $\alpha = 0$, speakers choose randomly and as $\alpha \rightarrow \infty$, they greedily choose the highest probability alternative. In our simulations below, we treat α as a free parameter and fit it to the data.

To instantiate our signaling game with a tractable message set M, in our studies we adopt a food-review paradigm: we assume that speakers and listeners are trying to communicate the number of stars in an online restaurant review (where $m \in \{1,2,3,4,5\}$). We then use experiments to measure three components of the model. First, to measure literal semantics $p_{L_0}(m \mid u)$ (we ask experiment participants to judge whether a message is compatible with a particular meaning (Experiment 1). Second, to generate a set of plausible alternative messages in U, we elicit alternatives directly (Experiment 2). Lastly, to obtain human L_1 pragmatic judgments, we ask participants to interpret a speaker's utterances (Experiment 3).

Experiment 1: Literal listener task

Experiment 1 was conducted to approximate literal listener semantic distributions $p_{L_0}(m \mid u)$ for five pairs of scalar items taken from Tiel (2014). We include three conditions in Experiment 1, corresponding to the sets of alternatives within a scale that participants were presented with: two-alternatives ("entailment"), four-alternatives, and five-alternatives. The two-alternatives entailment condition makes a test of the hypothesis that the two members of the classic Horn scale (Horn, 1972) are the only alternatives necessary to predict the strength of listeners' pragmatic inference. The four- and five-alternatives conditions then add successively more alternatives to test whether including a larger number of alternatives

Figure 2: (Left) A trial from Experiment 1 (literal listener) with the target scalar 'liked.' (Right) A trial from Experiment 3 (pragmatic listener) with the target scalar 'liked.'

will increase model fit.[1] A secondary goal of Experiments 1a,b,c tests whether the set of alternatives queried during literal semantic elicitation impacts compatibility judgments. (If it does we should see differences in compatibility judgments for shared items between experiments.) We address this possibility in the results section.

Methods

Participants In each condition we recruited 30 participants from Amazon Mechanical Turk (AMT). In the two-alternative condition, 16 participants were excluded [2] for either failing to pass two training trails or were not native English speakers, leaving a total sample of 14 participants. In the four-alternative condition, 7 participants were excluded for either failing to pass two training trials or were not native English speakers, leaving a total sample of 23 participants. In the five-alternative condition, 3 participants were excluded for either failing to pass two training trials or were not native English speakers, leaving a total sample of 27.

Design and procedure Figure 2, left, shows the experimental setup. Participants were presented with a target scalar item and a star rating (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 thought the food was ____?" where a target scalar was presented in the blank. Each participant saw all scalar item and star rating combinations for their particular condition, in a random order.

The two-alternatives condition included only the scalar pairs from Tiel (2014). The four-alternatives condition included the two scalar items plus the top two alternatives generated for each scalar family by participants in Experiment 2. The five-alternatives condition included the four previous items plus one more "neutral" item chosen from those alternatives generated in Experiment 2. See Figure 1 for the complete list of alternatives used in each condition.

Results and Discussion

Figure 4 plots literal listener $p_{L0}(m|u)$ scalar item distributions for the three conditions. Each row shows a unique scalar family with items ordered horizontally by valence. Several trends are visible. First, in each scale, the alternatives spanned the star scale, such that there were alternatives that were highly compatible with both the lowest and highest numbers of stars. Second, there was clear variability between scalar families. For example, compatibility judgments for the top items in the *memorable / unforgettable* scale were more similar than those for *good / excellent* or *liked / loved.* Finally, there was substantial consistency in ratings for items that were repeated across experiments, suggesting that this paradigm elicited stable judgments from participants.

To test our secondary hypothesis, that the different sets of scalar items queried might result in differences in compatibility judgments between experiments, we ran a mixed effects model. We regressed compatibility judgments on scale, number of stars and experiment (1a,b,c), with subject and word level random effects, which was the maximal structure that converged. Results indicate no significant differences between Experiment 1a and 1c, b =-0.05, Z = -0.53, $p = 0.59$ or between 1b and 1c, b = -0.04, Z = -0.52 $p = 0.6$ and the addition of our experiment predictor did not significantly improve model fit when compared to a model without the experiment variable using ANOVA $\chi^2(\ 2\ ) = 0.43$, $p = 0.81$.

Experiment 2: Alternative Elicitation

To elicit empirical alternatives for the scales we used in Experiment 1, we adopted a modified cloze task inspired by Experiment 3 of Tiel (2014).

Methods

Participants We recruited 30 workers on AMT. All participants were native English speakers and naive to the purpose of the experiment.

Design and procedure Participants were presented a target scalar item from our original entailment set (see Figure 1) embedded in a sentence such as, "In a recent restaurant review someone said they thought they the food was ____," with a target scalar presented in the blank. Participants were

[1] Note that alternatives in the four- and five-alternatives conditions were chosen on the basis of Experiment 2, which was run chronologically after the two-alternative condition; all literal listener experiments are grouped together for simplicity in reporting.

[2] The majority of respondent data excluded from Experiment 1a was caused by failure to pass training trials. We believe the task may have been too difficult for most respondents and made adjustments to the training trials in Experiments 1b,c

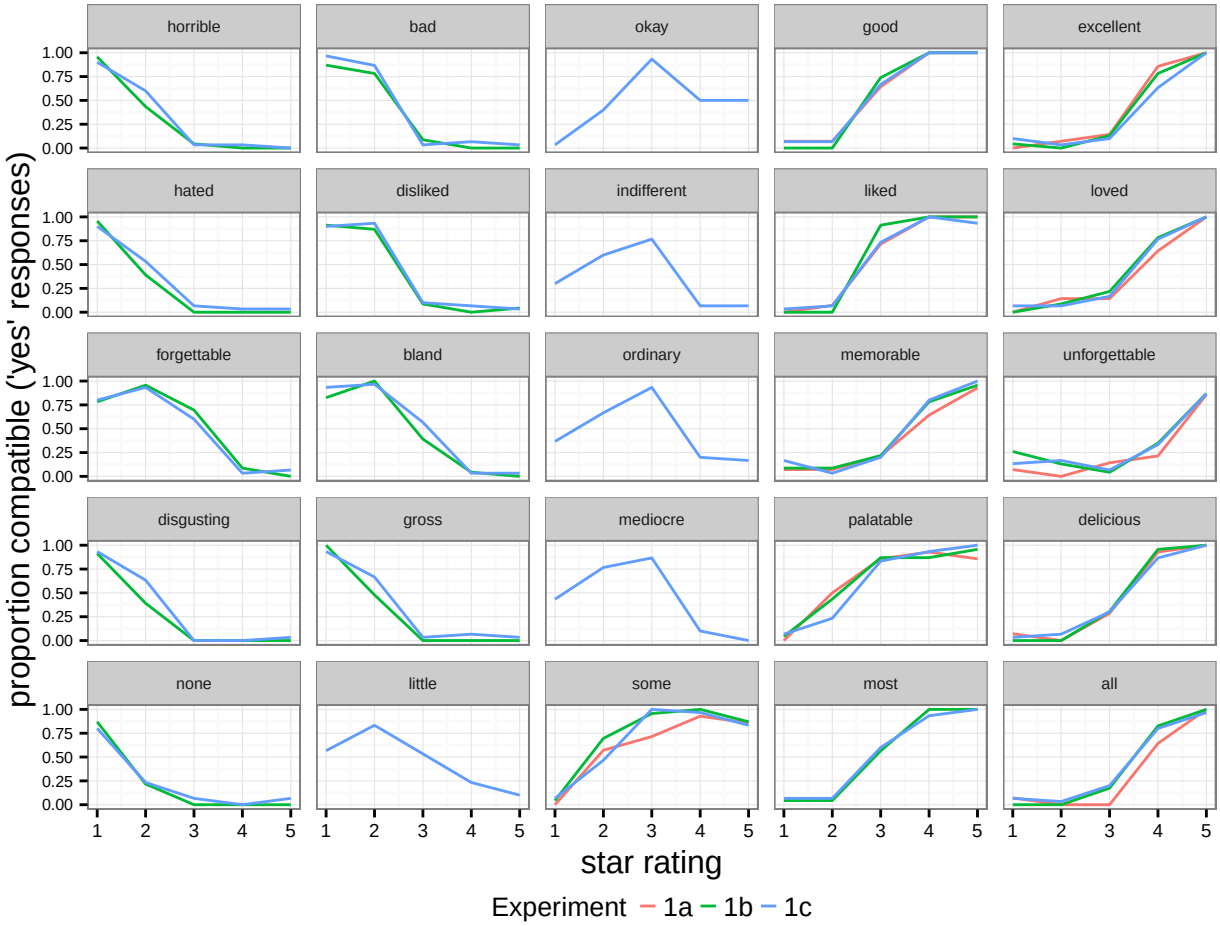

Figure 3: Literal listener judgments from Experiments 1a,b,c. Proportion of participants indicating compatibility (answering 'yes') is shown on the vertical axis, with the horizontal axis showing number of stars on which the utterance was judged. Rows are grouped by scale and items within rows are ordered by valence. Colors indicate the specific experiment (1a,b,c) with experiments including different numbers of items.

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 ____?" They were prompted to generate three unique alternatives.

Results and Discussion

Figure 5 shows an example alternative set for the scalar items *liked* and *loved* (combined). Alternative distributions for the other scalar pairs (e.g.. *good/excellent*, *memorable/unforgettable*) were similarly long-tailed.

Experiment 3: Pragmatic Listener

Experiment 3 was conducted to measure pragmatic judgments $p_{L_1}(m \mid u)$ for five pairs of scalar items taken from Tiel (2014). We include two conditions in Experiment 3. In the two-alternatives condition, participants made judgments for items included in the entailment scales. In the four-alternatives condition, participants made judgments for the entailment items and also the top two alternatives elicited for each scale in Experiment 2. Running two pragmatic judgment conditions allowed us to rule out the potential effects of having a larger set of alternatives during the pragmatic judgment elicitation and also provided two sets of human judgments to compare with model predictions[3]. A secondary goal of Experiment 3b tests whether the set of alternatives queried in Experiments 3a,b impacts pragmatic judgments. (If it does we should see differences in pragmatic judgments between experiments.) We address this possibility in the results section.

Participants

We recruited 100 participants from AMT, 50 for each condition. Data for 9 participants was excluded after participants either failed to pass two training trials or were non-native English speakers, leaving a total sample of 41 participants. In the four-alternative condition, data from 7 participants was

[3]Note that alternatives in the four-alternatives condition were chosen on the basis of Experiment 2, which was run chronologically after the two-alternatives condition; both pragmatic listener experiments are grouped together for simplicity in reporting.

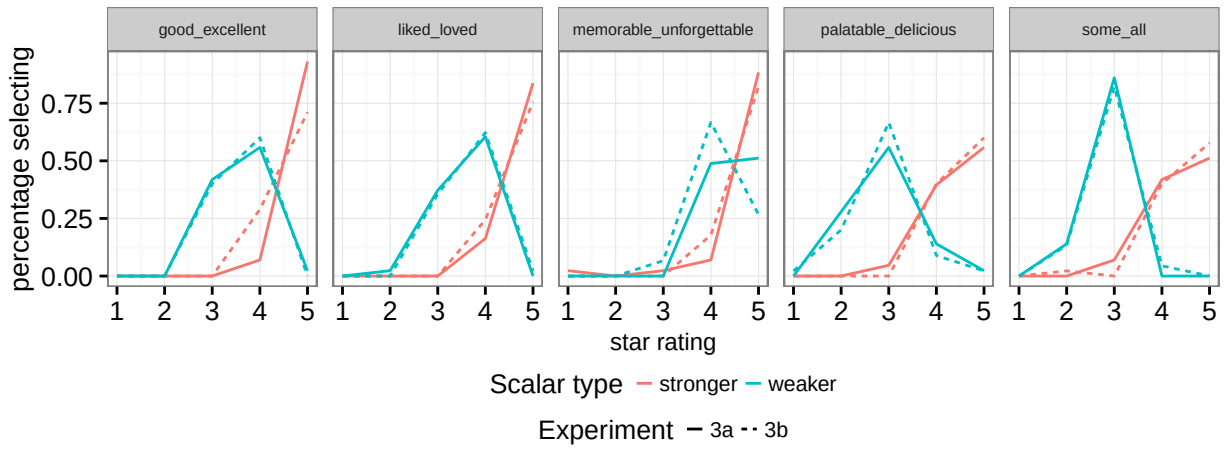

Figure 4: Pragmatic listener judgements for our entailment scalar items. Proportion of participants generating a star rating is shown on the vertical axis, with the horizontal axis showing number of stars on which the utterance was judged. Each line-type shows a different experiment, and colors indicate the stronger (ie excellent) or weaker (ie good) scalar items for each scale. Each panel shows one scalar pair. We are only showing scalar items used for model comparison (items used in Experiments 1a / 3a) for simplicity,

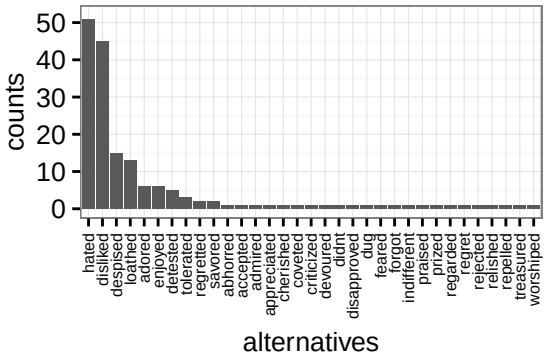

Figure 5: Combined counts for participant generated alternatives for 'liked loved' scale from Experiment 2. Participants were shown a target scalar (either liked or loved) and asked to give three alternatives.

excluded after participants either failed to pass two training trials or were not native English speakers, leaving a total sample of 43 participants.

Procedure

Participants were presented with a one-sentence prompt containing a target scalar item such as "Someone said they thought the food was _____." Participants were then asked to generate a star rating representing the rating they thought the reviewer likely gave. Each participant was presented with all scalar items in a random order. The experimental setup is shown in Figure 2, right.

Results and Discussion

Figure 4 plots pragmatic listener judgments distributions for "weak" / "strong" scalar pairs (e.g. *good/excellent*). We only include the original entailment pairs in this figure for simplicity. Several trends are visible. First, in each scale participants generated implicatures and were less likely to assign high star-ratings to weaker scalar terms, despite literal semantic compatibility. Second, the "size" of the difference between judgments between "strong" and "weak" scalar items varied by scale, consistent with results from Tiel (2014).

To test our secondary hypothesis, that the different sets of scalar items queried might result in differences in pragmatic judgments between experiments, we ran a mixed effects model. We regressed pragmatic judgments on scale and experiment (3a,b) with subject and word level random effects, which was the maximal structure that converged. Results indicate there were no significant differences between Experiment 3a and 3b, $b = 0.05$, $t(\ 150\ ) = 1.04$, $p = 0.3$ and the addition of our experiment predictor did not significantly improve model fit when compared to a model without the experiment variable using ANOVA $\chi^2(\ 1\ ) = 1.13$, $p = 0.29$.

Model Results

Using literal semantic data from Experiments 1a,b,c we conducted three simulations with our model. Each simulation used the specific literal semantic data to specify the scale representation available to our model.

Model	α	Exp. 3a	Exp. 3b
Two alts	9	0.54	0.57
Two alts + generic negative	6	0.62	0.66
Four alts	4	0.84	0.90
Five alts	4	0.86	0.90

Table 1: Model performance with fitted alpha levels. Model fit assessed through correlation with human judgments in our two Pragmatic listener experiments (3a,b)

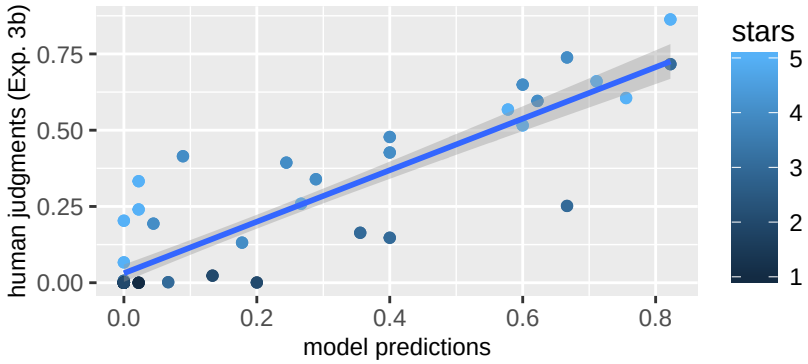

Figure 6: Model fit: model predictions are plotted on the x-axis. Human judgments from Exerpiment 3b are plotted on the y-axis. Coloration denotes star rating for which judgment was made.

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 judgments was significantly improved by the inclusion of alternatives beyond the typical "strong" and "weak" scalar items. The two-alternatives model contained only entailment items, which, under classic accounts, should be sufficient to generate implicature. Model performance with this limited alternative set was poor when compared to human judgments from Experiment 3a,b (row 1, Table **??**). Building off findings in Franke (2014), in which the best-fitting model assigned substantial weight to *none* as an alternative, we included a "generic" negative alternative that was only compatible with 1-star. Model fit for this modified entailment alternative set did improve, however overall correlation was still low (row 2, Table **??**). Model fit jumped substantially by adding the top-two alternatives for each scale (row 3, Table **??**). Finally, our five-alternatives model included a neutral alternative along with the original entailment pair and top two (row 4, Table **??**).

While model improvement seemed to be related to the number of alternatives present, we did not see substantial differences between our four- and five-alternative models. One possibility is that alternatives are differentially salient in context. Our current framework makes the simplifying assumption that all alternatives are equally salient. However, it may well be the case that while both *all* and *most* are *present* during pragmatic enrichment of *some*, *all* is more salient and therefore should carry greater weight in our model. Findings from Experiment 2, in which participants generated plausible alternatives, support this intuition, as some alternatives appeared more often than others. This distinction between binary *presence* and graded *salience* is an important one that deserves further investigation.

It is also likely that the precise set of alternatives present during implicature are domain dependent. Our current empirical paradigm elicited literal semantics, pragmatic judgments and plausible alternatives all within the restricted domain of restaurant reviews. Participant data may have differed greatly if we had instead asked them to respond in a movie review paradigm, which would have been acceptable for a number of our stimuli (e.g. *memorable/unforgettable*, *liked/loved*, etc.).

Considering the possibility that alternative sets may be domain specific even for scalar items may provide a way forward in uniting generalized and particularized implicatures. Findings from our investigation indicate that human literal semantic and pragmatic judgments are often graded and diverse, much like the set of plausible alternatives that give rise to pragmatic inferences.

Acknowledgements

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

References

Degen, J., & Tanenhaus, M. K. (2015). Processing scalar implicature: A constraint-based approach. *Cognitive Science*, *39*(4), 667–710.

Frank, M., & Goodman, N. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998.

Franke, M. (2014). Typical use of quantifiers: A probabilistic speaker model. In *Proceedings of the 36th annual conference of the cognitive science society* (pp. 487–492).

Goodman, N. D., & Stuhlmller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*(1), 173–184.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics* (Vol. 3). New York: Academic Press.

Horn, L. R. (1972). *On the semantic properties of logical operators.* (PhD thesis). University of California, Los Angeles.

Horn, L. R. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. *Meaning, Form, and Use in Context*, *42*.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. MIT Press.

Lewis, D. (1969). *Convention: A philosophical study*. John Wiley & Sons.

Tiel, B. van. (2014). Quantity matters: Implicatures, typicality, and truth.

Van Tiel, B., Van Miltenburg, E., Zevakhina, N., & Geurts, B. (2014). Scalar diversity. *Journal of Semantics*, ffu017.