The importance of full scale representations 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. In the canonical "some, but not all" scalar implicature, listeners enrich the meaning of "some" to "some, but not all" when interpreting a sentence such as "She ate some of the cookies." A Gricean account of this phenomena assumes the presence of "salient alternatives" with varying degrees of informatitivy. "Some," in the example above, is enriched to "some, but not all" in the presence of the stronger alternative "all." While intuitive, empirical evidence for the presence of such alternatives is limited. Our current work explores the role different scale representations may play in scalar implicature. Using a Bayesian model of implicature generation we simulate various scale representations that might be available to a listener, comparing model predictions to human judgment. Results indicate that implicature generation likely requires "full" scale representations, including negative valenced alternatives, rather than the typical entailment only scales assumed in most research to date.

Keywords: pragmatics; scalar implicature; bayesian modeling

Introduction

Humpty Dumpty Quote...

[Implicature as a case study for pragmatics]

Succesful communication regularly requires listeners to make inferences that go beyond the semantic content of a speaker's utterance. "Scalar implicatures" are a hallmark of such pragamtic 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." (citations) 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). Continuing with the prior example, within this framework a listener will infer that the speaker must have intended that Bob did not eat "all" the cookies because it would have been *underinformative* to use "some" when "all" could have been used.

[Scalar implicature in particular, what are the alternatives?]

The presence of "salient" alternatives is critical to standard accounts of scalar implicature (Grice). Previous research indicates implicature generation is affected by the set of alternatives available, both in adults (Degen, Franke) and is possibly the source of difficulty in computing implicatures for children (Barner). Degen found ... Similarily, Franke demonstrated... (the importance of a strong negative valenced term such as "none" adopting a Bayesian model of implicature generation. Model performance improved significantly when "none" was included...) Despite these signals, little empirical evidence points directly to the impact of scale representaion in scalar implicature. The current work investigates the role different scale representaitons play in scalar implicature.

[RSA as a way of studying implicature]

Measuring the degree to which alternatives may be present during implicature generation is a difficult task. Issues of introspection aside, simply querying a participant about the apparent saliency of particular alternatives is problematic because the alternatives are made salient in the query itself. To circumvent this issue, we adopted a computational model of implicature generation. Adopting a computational framework allowed us to simulate the effects of various scalar alternatives sets on implicature generation, comparing model predictions to human judgments.

[How RSA works]

Rational Speech-act theory (RSA) is a Bayesian framework for modeling scalar implicature. How it works... Previous work using RSA has accurately modeled human judgments in both ad-hoc (Frank & Goodman, 2012) and embedded implicature settings (Stuhmuller & Goodman ...)....

[Using the scalar diversity approach to get beyond the some/all implicature]

Recent experimental investigations of scalar implicature have focused almost exclusively on a small subset of possible scales, most notably focusing on "some/all". This lack "Scalar Diversity" (van Tiel, 2013) not only presents a problem in terms of generalization of findings, but within our current computational framework, presents a problem of data sparsity. We adopt a "Scalar Diversity" approach in our experimental design in order to incorporate multiple scalar items from a range of grammatical classes. In particular, we use a food review paradigm to quantify what would otherwise be ambiguous semantics as distirbutions over star-ratings.

[Outline of the rest of the paper]

In the following paper we present three experiments and three model simulations based on the experimental data. We used a food review paradigm in which linguistic 'meaning' for a given scalar item was quantified as a distribution over star-ratings. Quantifying both literal semantic content and pragmatic judgments in this way provided an important interface to RSA, especially for scalar items with ambigous literal semantics (items such as "memorable/unforgettable" or "palatable/delicious). 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. Alternatives generated in Experiment 2 suppl-

mented the original stimuli from Experiments 1a,b in Experiments 3a,b. Using the literal semantic data from Experiments 1a and 3a we modeled pragmatic enrichment using RSA and compared model predictions to human judgment data from Experiments 1b and 3b. Results indicated that model performance was significantly improved with the incorporation of negative valenced scalar items. Indeed, peak model performance occurs for simulations with the "fullest" scale representations. This result suggests that more "fully specified" scale representations may be active for human participants while making pragmatic judgments.

Experiments - Literal listener and Pragmatic Listener tasks

We used a food review paradigm to quantify literal semantics and assess pragmatic judgments for scalar items. Within this framework, 'meaning' for a given scalar term is quantified as a distribution over stars. Literal semantics for scalar items was assessed through a compatibility measure in Experiments 1a and 3a. Participants were presented a star-rating (betwen 1-5 stars) and asked "Do you think the person *loved* the food." We used a binary dependent measure "yes/no" to assess compatibility. To assess pragmatic judgments, participants were presented a sentence such as "Someone said they *loved* the food" and asked to select the number of stars they thought the reviewer likely gave. In this case, the dependent measure was a star-rating from 1-5 stars.

Experimental 1a,b: Entailment scales

Experiment 1a and 1b were conducted to approximate literal listener semantic distributions (study 1a) and assess pragamtic judgments (study 1b) for positive valenced scalar pairs (see appendix A for the complete list).

Participants

Particpants for both studies were recruited on Amazon Mechanical Turk and paid $0.20 for their participation. Thirty participants were recruited for Experiment 1a (Literal Listener task). Data for N particpants was thrown out after particpants either failed to pass two training trials or were not native English speakers, leaving a total sample of N.

Fifty particpants were recruited for Experiment 1b (Pragmatic Listener task). Data for N particpants was thrown out after particpants either failed to pass two training trials or were not native English speakers, leaving a total sample of N.

Procedure

On each trial of the literal listener task (study 1a) participants were presented with a star-rating (between 1-5 stars) and asked to judge the compatibility of the star-rating with a target scalar term. Compatibiltiy 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.

On each trial of the pragmatic listener task (study 1b), participants were presented with a one-sentence prompt containing a target scalar item such as "Someone said they ___ed the food" and were 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

Experiment 2: What are the alternatives?

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

Participants

Using Amazon's Mechanical Turk, N workers were paid $0.20 to particpate. 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 purposeof the experiment.

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

Experiment 3a,b: Full scales - incorporating negative alternatives

Experiment 3a,b was conducted to approximate literal listener semantic distributions (study a) and assess pragamtic judgments (study b) for both positive and negative valenced scalar items (see appendix b for the complete list of items used).

Participants

Particpants 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 particpants was thrown out after particpants either failed to pass two training trials or were not native English speakers, leaving a total sample of N.

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

Procedure

Results and Discussion

Model

Details

Model fitting and comparison

General Discussion

Acknowledgements

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

References

the role various scale representations might play in scalar implicature. In Experiment 1a,b we gather data for Entailment only scalar items such as "some/all," "good/excellent," etc. In Experiment 1a we quantify literal semantics for scalar items as compatibility distributions over star-ratings (Literal listener task). In Experiment 1b, we ask participants to make judgments about speaker intentions by generating star-ratings in response to reviews (Pragmatic listener task). In Experiment 2, we ask participants to generate plausible alternatives to the scalar items used in Experiments 1a,b. Experiments 3a,b are identical to 1a,b with the addition of scalar items from Experiment 2. In this we explore fuller scale descriptions, gathering both literal semantics and pragmatic judgments for fuller scale descriptions such as "none/some/most/all" or "horrible/bad/good/excellent."

In three experiments we Our "Scalar Diversity" approach involves a food review paradigm we employed throughout our three experiments. In both Literal Listener (1a, 3a) and Pragmatic Listener (1b, 3b) tasks, participants are shown one sentence prompts containing a target scalar term along with a star-rating. In individual trials, star-ratings were either pre-populated (in literal listener tasks) or used as the dependent variable (pragmatic listener tasks). This design allowed us to quantify literal semantics for individual scalar terms as compatibility distributions over star-ratings (literal listener tasks 1a and 3a). Similarily, we quantified pragmatic judgments by prompting participants with a target scalar term and asking them to generate a star-rating.