Modeling Scalar Diversity: Quantifying Ambiguous Scalar Semantics

Ben Peloquin
Stanford University

Abstract

Over the last decade, experimental and computational investigations of scalar implicature have focused on a few exemplar cases, such as <some, all>. Recent experimental work suggests a high degree of variability in implicature rates for different lexical scales. This amounts to a greater degree of "scalar diversity" than was previously assumed. We argue that the difficulty in addressing scalar diversity from a computational perspective stems from difficulty quantifying ambiguous scalar semantics for items such as <good, excellent> or <palatable, delicious>. We present two possible methodologies for quantifying ambiguous scalar semantics using an experimental paradigm and corpus based approach with application to rational speech-act theory.

Keywords: Natural language pragmatics; Bayesian model; Scalar implicature

Note: This work is ongoing with only preliminary data analysis completed. Charts do not currently contain error bars, but are meant to communicate initial findings and illustrate potential productivity of the approaches accounted for in this paper.

Introduction

A listener's capacity to infer meaning beyond the lexical content of an utterance has been a major topic of interest in both philosophy of language and psycholinguistics for nearly half a decade and falls under the domain of "natural language pragmatics." In particular, the field has been shaped by Paul Grice's seminal work, "Logic and Conversation" (1975). To date, the Gricean account of *pragmatic implicature* (a Gricean term for the ability to infer extra-linguistic meaning in communication) has proved to be productive in both experimental (Huang & Snedeker, 2009, Papafragou & Musolino 2003, and many more) and computational settings (Frank & Goodman 2012, Goodman & Stuhlmuller 2014, Lassiter & Goodman, 2014).

This productivity stems primarily from Grice's identification of the generating factors behind pragmatic implicature. These factors are amenable to various experimental paradigms as well as computational models involving Bayesian inference and information theoretic notions around informativity. However, recent experimental work examining *scalar implicature* (pragmatic inference between scalar items) by van Tiel, et. al (2014), highlights an important observation for both experimental and computational investigations of pragmatic inference. In the domain of scalar implicature, this observation can be summarized as follows – not all implicatures are created equal.

The current paper is concerned with the steps we might take to capture the idea of non-uniformity of scalar implicature in a probabilistic model. To motivate this issue, we first familiarize the reader with the Gricean framework in relation to scalar implicature and introduce van Tiel's concept of *scalar diversity*. We then give a brief introduction to Rational Speech Act theory, highlighting the fundamental difficulty in capturing ambiguous scalar semantics, which, we propose, is necessary for modeling diversity in implicature rates between different lexical scales. We present two approaches for doing this, one experimental and one corpus based, and discuss the implications moving forward.

The Gricean account

In "Logic and Conversation" (1975), Grice presents an intuitive framework for examining pragmatic inference. At its core, the Gricean account involves a *Cooperative Principle* and four maxims. These are best understood as guides for examining pragmatic meaning, rather than rigorous scientific theory. The Cooperative Principle, "make your conversation contribution such as is required… by the accepted purpose or direction of the talk exchange in which you are engaged" (Grice, 1975), serves as context for rational communication between interlocutors. Within this context, a listener may modulate meaning interpretation to maintain the assumption of speaker cooperativity.

In addition to the Cooperative Principle, Grice presents the maxims of *Quantity* (be as informative as possible, but no more), *Quality* (do not say what you believe to be false), *Relation* (be relevant) and *Manner* (avoid obscurity and ambiguity, be brief). A *conversational implicature* (another Gricean term for inference in conversation) is best understood as the inference a listener is compelled to make in attempting to maintain that the speaker is abiding by the Cooperative Principle, given the interaction of potentially competing maxims.

An example with scalar implicature

Imagine a scenario in which Bob and Alice live together. Bob recently baked a batch of cookies. Knowing Alice is a cookie lover who could finish off a batch en passant, Bob tells her "you can eat some of the cookies." Does Bob mean Alice can eat *all* of the cookies? While the logical meaning of "some" is "at least one," and therefore compatible with the scenario in which Alice eats all the cookies, most listeners infer that Bob means to communicate that Alice may *not* eat all of the cookies. This is an example of scalar implicature. Within the Gricean framework, scalar implicature arises when a speaker uses a word that evokes a lexical scale (such as <some, all>). Each scale member represents a salient alternative to the others, and the terms are conceptually ordered by informativeness (van Tiel, 2014). A speaker who uses a less than maximally informative scalar term (such as “some” in the context of “all”) often implies that they do not believe the more informative scalar term was appropriate. Under the Gricean account, this interaction between a given scalar’s informativeness and its alternatives, represents an interaction of the Quantity and Quality maxims.

We make this intuition explicit by again referring to the example with Bob, Alice and the cookies. If we assume Bob is cooperative and obeying the maxim of Quantity, we’d expect him to use the scalar term “all” if his intended meaning was for Alice to eat as many cookies as she pleased with no upper bound. Under the Gricean framework, we reason that Bob must not have been able to use the stronger scalar term “all” because this would have clashed with the maxim of Quality (be truthful), if in fact Bob did not intend for Alice to eat all of the cookies. Therefore, we reason that Bob must have chosen the maximally informative scalar “some”, *given* that he does not intend for Alice to eat all of the cookies. Through this iterated reasoning we arrive at the conclusion that Bob’s intended meaning must be that Alice may eat *some, but not all* of the cookies.

Scalar diversity and the uniformity assumption

Over the last decade of research in experimental pragmatics, the “some, but not all” implicature has surfaced as the overwhelming favorite in experimental work across design paradigms. Recent work by van Tiel and colleagues (Scalar Diversity, 2014) has exposed this reality and called attention to the implicit assumption therein. The “Uniformity Assumption” states “observations about the behavior of a particular lexical scale can typically be generalized to the whole family of lexical scales” (p. 4). This assumption was tested by van Tiel and colleagues using an inference paradigm, which assessed implicature rates for over 30 different scalar pairs. Results indicated a high degree of variability in implicature rates for different lexical scales, evidence against uniformity and for the concept of *scalar diversity* (van Tiel, 2014).

This observation raises questions for future experimental work and has important implications for how computational models of scalar implicature treat different scalar items. As we build more sophisticated models of pragmatic inference and scalar implicature in particular, we should address the idea of scalar diversity by developing formulizations sensitive to between-lexical-scale variability. That is, we should account for the observation that “not all implicatures are created equal” in our models. To motivate this intuition we first present a brief introduction to RSA. Following this introduction we argue that a critical barrier to modeling scalar diversity stems from the inherent difficulty of quantifying ambiguous scalar semantics. We present two possible methodologies to overcome this obstacle.

Formalizing the Gricean account: rational speech-act theory

Rational speech-act (RSA) theory formalizes Gricean intuitions in a computational model. In RSA, we represent the probability of a listener interpreting a particular meaning m given a speaker’s utterance u as a simple application of Bayes’ rule, where the posterior is proportional to the speaker likelihood $P_{speaker}$ multiplied by the listener’s prior beliefs about the world state $P(m)$:

$$P_{listener}(m|u) \propto P_{speaker}(u|m) \times P(m)$$

Under RSA we assume that a speaker chooses an utterance according to Bayesian decision theory - choosing words according to their expected utility such that

$$P_{speaker}(u|m) \propto e^{\alpha U(u;m)}$$

where the utility function $U(u;m)$ quantifies the value of saying u given that the intended meaning (or state) is m. Crucially, utility is related the amount of information that a literal listener $P_{listener}$ would not yet know about m after hearing u. This quantity is equivalent to the negative surprisal (Goodman & Stuhlmuller, 2012) and represents a challenge in extending RSA to scalar terms with ambiguous scalar semantics.

$$U(u;m) = \ln(P_{literal}(m|u)) - C(u)$$

The basic form of this model has successfully captured the nature of iterated social reasoning and causal learning (Goodman, Baker & Tenenbaum, 2009), social reasoning in language games (Goodman & Frank, 2012) as well as scalar implicature and its interaction with knowledge about speaker epistemic state (Goodman & Stuhlmuller, 2013). Overall, RSA approximates the inference a pragmatic listener would make given a speaker she assumes to be rational and having the goal of being informative (Goodman & Stuhlmuller, 2012).

Scalar semantics and informatitivy

Consider how we might apply RSA to the example of Bob and Alice. Let’s say that Bob baked five cookies and tells Alice she can eat “some” of the cookies. In order to capture the implicature “some, but not all” we need to quantify the logical meanings of “some” and “all.” In this case, both “some” and “all” quantify over sets so we can decompose the their semantics to the possible world states they are compatible with (the states in which they are true). If n is the number of cookies Bob thinks it’s OK for Alice to eat, then “some” will be true in any circumstance in which Bob’s meaning is that Alice may have more than 0 cookies:

SOME is true in any state in which $n > 0$

Alternatively, “all” is only true in the instance in which Bob thinks it’s OK for Alice to eat all five of the cookies:

ALL is true in state $n == 5$

We could imagine the existence of another alternative “none” which could be formalized as:

NONE is true in state n == 0

Given these quantities we can now approximate our literal listener informativity as the compatibility of a given meaning (say Bob intended that it was OK for Alice to have n==2 cookies) with the scalar term uttered. Since “some” is compatible with each of n = {1, 2, 3, 4} then:

$$P_{literal}(n == 2|"some") == 1/4$$

$$P_{literal}(n == 2|"all") == 0$$

$$P_{literal}(n == 2|"none") == 0$$

Accounting for ambiguous scalar semantics

Since we can quantify the semantics for “some” and “all” quite literally over a set of items, such as cookies, it is trivial to integrate into our current RSA framework. But how would we capture the semantics of scalar terms like <good, excellent>, <liked, loved>, <memorable, unforgettable> or even <palatable, delicious>? In “Processing Scalar Implicature: A Constraint-Based Approach,” Judith Degen and Michael Tanenhaus present a probabilistic Constraint Based framework which assesses implicature generation through a *naturalness* measure. Degen and Tanenhaus argue that naturalness and the availability of alternatives play a central role in computing a “some, but not all” implicature. Naturalness is measured using a 7pt Likert scale.

By assessing the degree of naturalness of a given utterance Degen and Tanenhaus introduce an important conceptual framework for measuring implicature – the idea that it can be graded or captured as a distribution over quantities. In the following section we propose two methodologies, each of which attempt to capture scalar semantics as distributions over a quantity of star ratings. We do so using both experimental data as well as corpus data using the Yelp academic data set.

Star rating paradigm

In order to quantify otherwise ambiguous lexical semantics for scalar terms such as <palatable, delicious>, we developed a star-rating paradigm to assess the relative compatibility of scalar term with a given star rating. The rational was that scalar terms such as <good, excellent> should display different distributional properties over star ratings such that the stronger scalar term “excellent” should be more compatible with higher star ratings, while the weaker scalar term “good” should have broader compatibility across multiple star ratings (as in the case with “some’s” compatibility across multiple cookie world states in our Bob/Alice example).

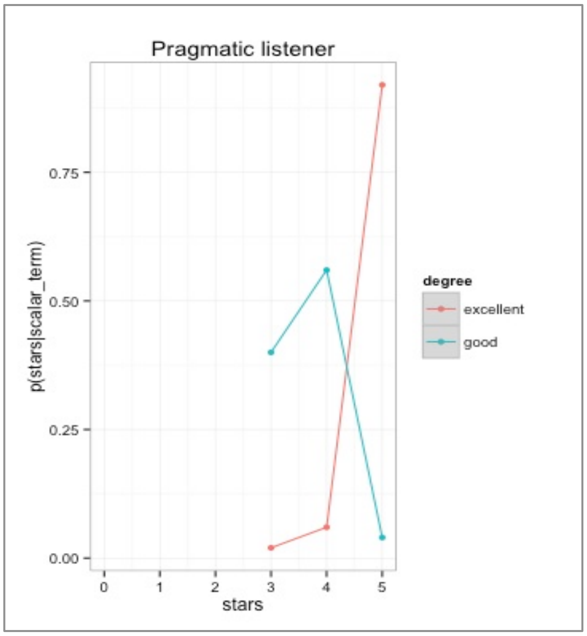

Figure 1. Pragmatic listener judgments: Good/Excellent

Experiment 1: Pragmatic listener judgments

Experiment 1 was conducted to determine pragmatic listener judgments ($P_{listener}(m|u)$). These judgments served as our benchmark for comparison with model predictions.

Method

Participants Using Amazon’s Mechanical Turk, 50 workers were recruited to participate in our study. All were native English speakers.

Procedure and materials Participants were told that they were participating in a study to try to better understand what people meant by their star ratings. In each trial, participants were shown a sentence with a target scalar such as “someone said the food was *good*” and asked to choose the star rating they thought the reviewer gave (see Fig. 1 for an example with <good, excellent>). Presentation of the target scalar terms was randomized and participants were presented each scalar term from the pairs <some, all>, <liked, loved>, <good, excellent>, <memorable, unforgettable> and <palatable, delicious>.

Results

This paradigm appears to have elicited substantial implicatures between competing scalar terms. We can observe this as the delta between the “high” (stronger scalar term) and “low” (weaker scalar term) for 4 and 5 star ratings in Fig. 1. Interestingly, in most cases participants never even selected ratings below 3 stars. Presumably this is because we only tested positive valence scalars, which

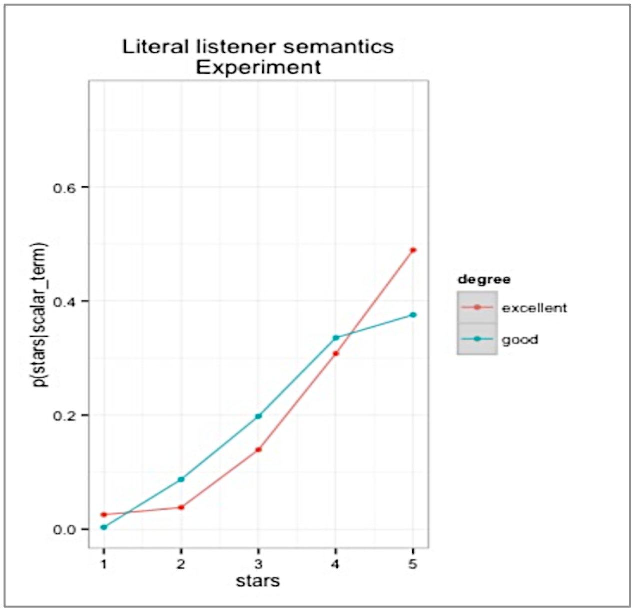

Figure 2. Experiment: Good/Excellent literal listener semantics

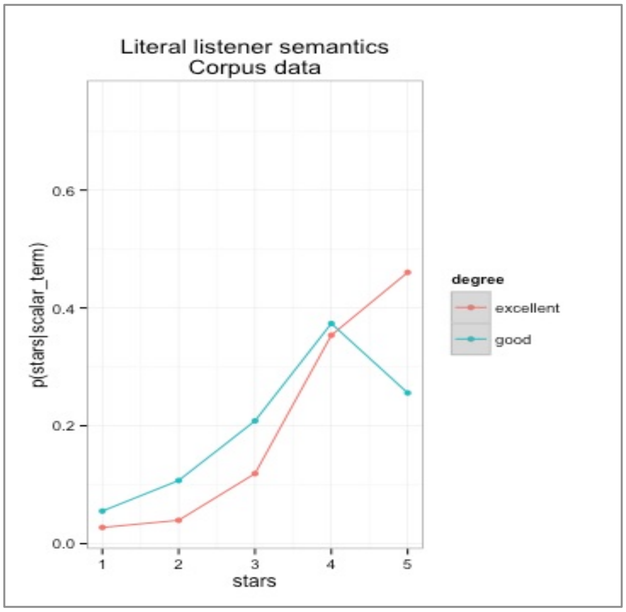

Figure 3. Corpus data: Good/Excellent literal listener semantics

intuitively have higher compatibility with higher star ratings.

Experiment 2: Literal listener semantics

Experiment 2 was conducted to approximate scalar semantics as a compatibility distribution over star ratings – in other words, our literal listener semantics ($P_{literal}(m|u)$).

Method

Participants Using Amazon's Mechanical Turk, 30 workers were recruited to participate in our study. All were native English speakers.

Procedure and materials Participants were told that they were participating in a study to try to better understand what people meant by their star ratings. Participants were first presented with a star rating (such as 3 out of 5 stars) and asked if they agreed that the reviewer felt ____ about the food, where ____ was filled by a target scalar. For example, in one trial participants would have been shown 5 stars and asked "how much do you agree that the person thought the food was *good*?" We used a 5pt likert scale to assess scalar compatibility with a given star rating. Each participant was shown all possible combinations of scalar target words with star ratings. So a single participant would be asked about the compatibility of the scalar "good" with a star rating of 1 through 5 stars. The ordering of scalar terms and star ratings were randomized throughout the experiment.

Results

Given that we assessed the graded nature of a target scalar's compatibility with a given star rating using a Likert scale, we needed to transform responses and then normalize across all possible star ratings to get a distribution over ratings. We did this by simply subtracting one from each star rating judgment, dividing that quantity by four (so that it sits in the interval [0, 1]), averaging this quantity across participants and then normalizing across star ratings. This transformation results in a distribution for each scalar term across star ratings. Fig. 2 displays the distribution for the scalar pair <good, excellent>. While this paradigm appears to capture the idea of a graded compatibility over star ratings, it is clear that the differentiation is not robust across all stars and seems to be most differentiated for the highest 5 star rating.

Experiment 3: Priors

Experiment 3 was conducted to assess prior beliefs participants had about star ratings.

Method

Participants Using Amazon's Mechanical Turk, 50 workers were recruited to participate in our study. All were native English speakers.

Procedure and materials Participants were told that they were participating in a study to try to better understand what people meant by their star ratings. Participants were presented with six possible scenarios in which they were asked give a star rating given no additional information. For example, in one trial participants might have been shown a sentence like "John went out to a restaurant with a friend. Without knowing anything about the food he ate, how many stars do you think he gave the restaurant?" Participants were then asked to make their best guess about the star rating (out of 5 stars). Names and the order of scenarios were randomized throughout the experiment.

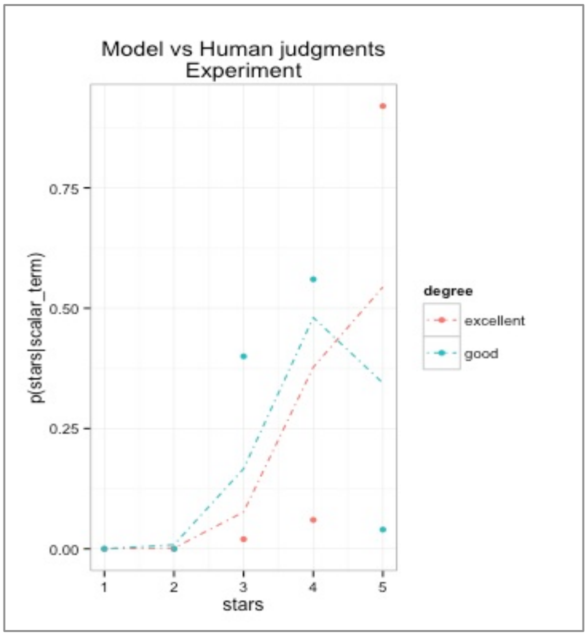

Figure. 4 Model comparison – Experimentally derived semantics

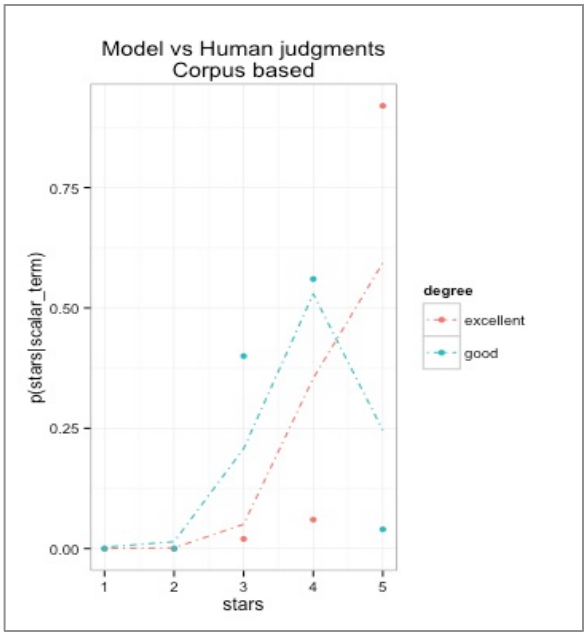

Figure 5 Model comparison – Corpus based semantics

Results

Responses tended to toward 3 and 4 stars and were fairly uniform across all contexts. We will make little mention of the priors from here on out. As it turns out, prior distributions varied to significant degree between the experimental and corpus based approaches. Because of this discrepancy we use uniform priors in the model. In addition, our method of introducing a context and asking for a star rating may have been an artificial and strange task for participants.

Corpus based approach: Yelp data set

The Yelp academic data set includes over 100,000 reviews for businesses across the U.S. For a portion of the data set, "Review Objects" contain both textual consumer review data as well as meta-data containing the star ratings associate with a given review. We parsed 10,000 Yelp review objects for occurrences of our target scalar items <some, all>, <good, excellent>, <liked, loved>, <memorable, unforgettable>, and <palatable, delicious>. We approximated scalar semantics as relative counts over star ratings. We also collected priors as simply a distribution over star-ratings regardless of occurrence of a scalar item. Fig. 3 displays the literal listener semantics for the scalar pair <good, excellent> derived from our corpus based approach.

Model performance

How does RSA perform when we approximate literal listener semantics as compatibility distributions over star ratings? We observed that implicature rates in our pragmatic listener condition (Exp 1.) were very robust (Fig. 1) while the scalar semantic distributions appeared to make little differentiation between scalar items (Fig. 2 and 3). As it turns out, RSA, in the basic form we have used here, using uniform priors and setting alpha to 2, does qualitatively predict implicature for all our scalar terms, but not nearly to at the effect size we observed in human judgments (Fig. 4 and 5). When we look at aggregate performance across scalar items, model predictions for both experimentally based and corpus based literal listener semantics tend to correlate at roughly $r = .69$ and $r = .71$ for each methodology, respectively (see Fig. 7 and 8)

Discussion and next steps

We have presented two plausible methodologies for estimating ambiguous scalar semantics with applications to rational speech-act theory. Both approaches captured the intuition behind graded semantic compatibility over states (or meanings). However, model predictions fell short of human judgments in both cases.

Adjustments moving forward

The lack of a robust implicature effect from the model may have stemmed from a number of dimensions from both the experimental paradigm and the corpus based approach. I highlight two possible explanations below.

Ambiguity surrounding the QUD Our current paradigm attempted to restrict participant judgments in both the pragmatic listener condition (Exp. 1) and the literal listener condition (Exp. 2) to be about food. However, ratings are often informed by highly subjective judgments in which reviewers take a variety of information into account, much of which may have little to do with the quality of the food. When we attempted to restrict participant responses to the food-only domain, this may seemed a bit artificial, possibility introducing a degree of ambiguity in terms of the question we being asked of them. We could attempt to address this by having participants respond within domains with a higher level of objectivity. For example, instead asking about food, we could present our scalar terms in the context of product reviews or business that have less subjective aspects such as a shoe repair shop or tailor.

Avoiding decontextualized corpus counts for scalar items

In terms of adjustments for the corpus based methodology there are a number of critical flaws with the current approach. First and foremost, simply recording scalar frequencies over star ratings without considering context (for example negation) is not an accurate depiction of a particular scalar's compatibility with a star rating. Future corpus based approaches should attempt to make use of language models, such as n-gram models which can incorporate contextual word information. Additionally, due to time constraints, our preliminary Yelp data set included some non-restaurant businesses. Ideally, we would restrict the reviews we parse to a specific domain (such as restaurants).

Acknowledgments

Thanks to Michael Frank and Paul Ganir for their help and suggestions throughout this project.

References

Degen, J. & Tanenhaus, M. K. (2014). Processing Scalar Implicature: A Constraint-Based Approach. *Cognitive Science, (1-44)*

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

Goodman, N. D., Baker, C. L. & Tenenbaum, J.B, (2009). Cause and Intent: Social Reasoning in Causal Learning. *Cognitive Science*.

Goodman, N. D. & Stuhlmuller, A. (2013). Knowledge and Implicature: Modeling Language Understanding as Social Cognition. *Cognitive Science*. 5 (173 – 184).

Goodman, N. D. & Tenenbaum, J. B (electronic). Probabilistic Models of Cognition. Retrieved <June 5th, 2014> from http://probmods.org

Grice, H. P.(1975). Logic and conversation. In Peter Cole & Jerry Morgan (eds.), *Syntax and semantics,* vol. 3: Speech Acts, 43-58, New York, Academic Press.

Huang, Y. T. & J. Snedeker. (2009). On- line interpretation of scalar quantifiers: insight into the semantics-pragmatics interface. *Cognitive Psychology* 58: 376–415.

Lassiter, D. & Goodman, N. D. (2015). Adjectival vagueness in a Bayesian model of interpretation. *Synthese*.

Papafragou, A. & J. Musolino. (2003), 'Scalar implicatures: experiments at the semantics-pragmatics interface'. *Cognition* 78:253–82.

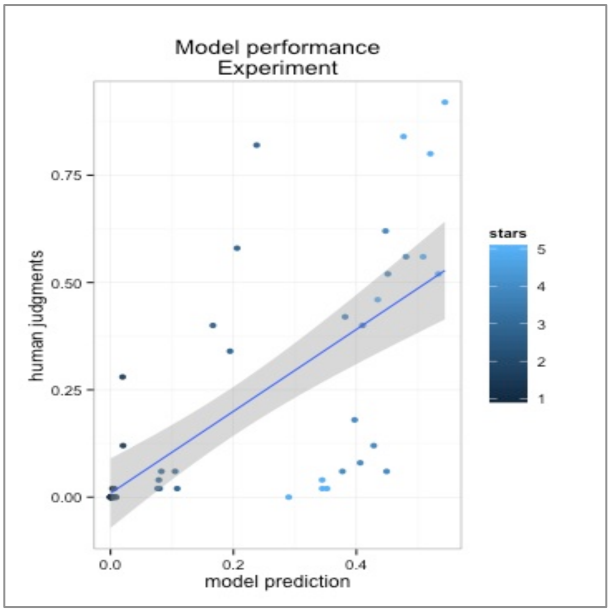

Figure 7. Model correlation – Experimentally derived semantics

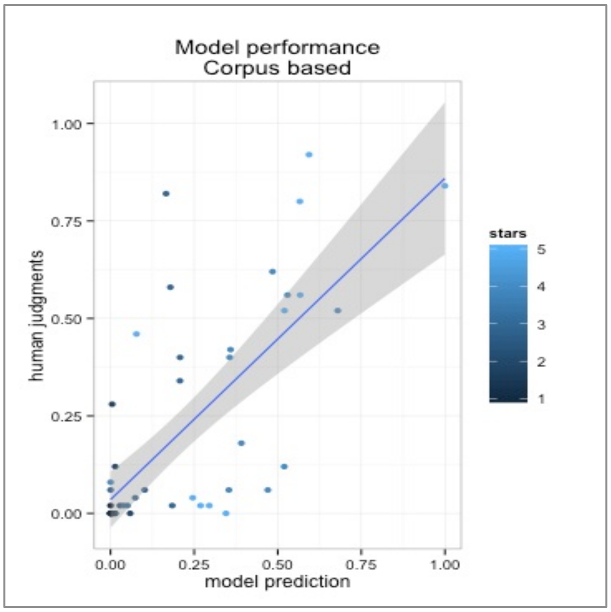

Figure 8. Model comparison – Corpus based semantics

van Tiel, B., et. al (2014). Scalar Diversity, *Journal of Semantics.* 0, (1-39)