Ad-hoc scalar implicature in adults and children

Alex Stiller
astiller@stanford.edu
Symbolic Systems Program
Stanford University

Noah D. Goodman
ngoodman@stanford.edu
Department of Psychology
Stanford University

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

Abstract

Linguistic communication relies on pragmatic implicatures such as the inference that if "some students passed the test," not all did. Yet young children perform poorly on tests of implicature, especially scalar implicatures using "some" and "all," until quite late in development. We investigate the origins of scalar implicature using tasks in which the scale arises from real-world context rather than conventional contrasts between lexical items. Experiment 1 shows that these ad-hoc implicatures are easy for preschool children, suggesting that children have an early competence at pragmatic inference, and that failures in standard scalar implicature tasks are due instead to problems contrasting lexical items. Experiments 2 and 3 compare a Gricean, *counterfactual* account of implicature with a *linguistic alternatives* account and find that neither predicts effects of contextual informativeness. We conclude that an account of pragmatic implicature must integrate world knowledge, linguistic structure, and social reasoning.

Keywords: Scalar implicature; pragmatics; language acquisition.

Introduction

Sometimes the absence of a description says just as much as its presence. A professor who says "some students passed the test" implies that some students failed—if all had passed, a cooperative speaker would have made the stronger statement "all students passed." *Scalar implicature* refers to the conversational shorthand of using weak terms to imply the negation of stronger ones that lie along the same "scale." In this paper we investigate the origins of scalar implicature, and the nature of scales, by investigating a spectrum of tasks that are logically equivalent to conventional scalar implicature but in which the scale arises (or fails to arise) from the real-world context rather than the lexical items—*ad-hoc implicatures*.

Implicatures surface in a variety of contexts beyond the case of quantifiers, including modal operators such as *"might"* and *"must"* (Noveck, 2001), inclusive and exclusive disjunction (Braine & Rumain, 1981), and numerals (Barner & Bachrach, 2010). A wide variety of theoretical frameworks have been proposed to explain implicature, with the two most influential being (1) Gricean approaches that we will collectively call the *counterfactual* theories (Grice, 1975; Levinson, 2000) and (2) views based on grammatically computed *linguistic alternatives* (Fox, 2007; Chierchia, Fox, & Spector, 2008).

Grice (1975) offers two maxims from which scalar implicatures are meant to follow: make your contribution as informative as is required, and do not make your contribution more informative than is required. From these it follows that any alternative statement which is more informative than the spoken statement must be false—because the speaker could have said that statement had it been true. Under this analysis, the relevant scale arises from the logical structure of the possible statements that could have, counterfactually, been uttered. Although neo-Gricean accounts have modified some parts of this basic inferential mechanism, the general predictions remain (Levinson, 2000). In response to apparent over-prediction of implicatures by the counterfactual theory, the linguistic-alternatives theory claims that implicatures arise by a process in which a statement is strengthened by negating the alternative statements—where the alternatives, and hence scales, derive from the lexical and grammatical structure of language; importantly, more complex statements are not taken to be alternatives (Fox, 2007; Chierchia et al., 2008).

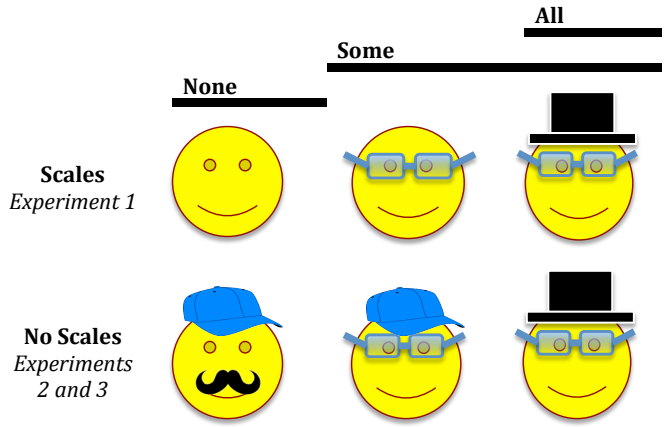

Figure 1: Example stimuli from our ad-hoc scalar implicature task. The utterance "My friend has glasses" receives different interpretations when the context given to the listener is Row 1 versus Row 2. Each has a similar logical structure to the conventional some-not-all implicature (top).

Consider the three situations shown schematically in Figure 1. At the top, the word "all" is logically stronger than the word "some", though some applies whenever all does. There is thus a natural scale of informativeness set up by the conventional semantic content of the words. In contrast, the feature words "glasses" and "top hat" have no conventional ordering, but in the context of the three faces in the middle row ("scales" condition), top hat is similarly stronger than glasses, though glasses applies to any object that top hat does. If a speaker says "the one with glasses" we may draw the implicature that she means the middle face (an intuition which we test in Experiment 1)—the situation itself seems to set up

the scale from which we can draw an implicature. However, this intuition is significantly weaker for the bottom row ("no scales" condition), despite an identical logical structure (an intuition we test in Experiment 2).

The acquisition of scalar implicature is late: children fail overwhelmingly at scalar pragmatic tasks where adults succeed (Papafragou & Musolino, 2003). Smith (1980), in one of the earliest acquisitional investigations of scalar pragmatic abilities, found that children with syntactic mastery of quantifiers such as "some" and "none" still failed to make the some/not-all implicature. Noveck (2001) found that 87% of children accepted statements such as "Some elephants have trunks" whereas only 41% of adults did. Huang and Snedeker (2009) replicated these findings, observing that children between ages five and nine construe weak scalar statements logically, while adults interpret such statements pragmatically. We use the ad-hoc implicature setting to ask whether developmental delays in success at scalar implicature tasks reflect inability to draw pragmatic inferences at all, or inability to access the scales inherent in the conventional semantics of some words.

Thus, the goal of the current experiments was to explore two related questions. The first was whether children younger than those tested in standard linguistic scalar implicature tasks would be able to succeed in ad-hoc implicatures (Experiment 1). The second was to explore the roots of "scales" in pragmatic inference, and the ability of either simple counterfactual or simple linguistic-alternatives theories to explain scalar implicature in general (Experiments 2 and 3).

Taken together, the results of our studies suggest that the inferential mechanisms underlying scalar implicature may be present earlier in development than previously assumed, and that the scales involved in implicature derive from world-knowledge, such as the base-rates of different properties, rather than logical structure of the context or conventional linguistic knowledge alone. Our data rule out the simplest version of both the Gricean counterfactual theory and the linguistic alternatives theory. Instead, they point the way towards an account in which linguistic and social factors are integrated probabilistically with world knowledge.

Experiment 1

Experiment 1 compares the performance of adults and children at an ad-hoc implicature task in which a pragmatic inference derives from an explicit context. We constructed a paradigm in which participants heard a sentence whose literal meaning was ambiguous between two referents. Though no conventional scale existed among the possible descriptions of these objects (e.g. "has glasses", "has top hat"), they varied along a contextually salient scale (Figure 1, middle). If participants were competent at pragmatic inference we expect them to succeed on this task, even though they may fail at an equivalent task in which the scale depends on lexical knowledge.

Methods

Participants Data were collected from 12 3–4 year-olds (M=3;6) and 12 4–5 year-olds (M = 4;5) at Stanford's Bing Nursery School. Twenty-four adult participants were recruited via Amazon's Mechanical Turk web-based crowdsourcing platform.

Stimuli We constructed pictorial stimuli for four trials—faces, houses, pasta, and beds. For each inference trial, two properties varied (e.g., presence of glasses and top hat for the "face" trial). In each trial, three objects were presented—one with neither feature (Distractor), one with exactly one feature (One-feature), and one with both features (Two-feature). Positions of the three objects were counterbalanced over six orders. Example stimuli are shown in Figure 1. Two unambiguous filler trials were constructed: in these, participants were asked to pick a car of a particular color and a fruit of a particular type.

Procedures In each trial (inference and filler), preschoolers were presented with three alternatives shown on laminated cards; adults performed a parallel task, picking alternatives by clicking on corresponding radio buttons in a webpage. The cover story involved a puppet, Furble, who asked the participants to help identify various people and objects. For example, in the "house" trial Furble said "My house has a flower outside. Can you show me my house?" The adult version involved the same script, with a picture of Furble substituting for the puppet. On each experimental trial Furble used a description that could apply to either the One-feature or the Two-feature object (e.g. "My friend has glasses" in Figure 1, middle). Adults were informed that the task was designed for children.

Results and Discussion

All three groups performed above chance, selecting the One-feature item for which a more informative description was not available (rather than the Two-feature item, with an alternative, unique description or the Distractor). Means and confidence intervals are shown in Figure 2. We analyzed these results using a logistic mixed-effects model (Gelman & Hill, 2007), predicting correct performance as a function of age group, with crossed random effects of participant and item. Coefficient estimates from this model are shown in Table 1. Adults were reliably more accurate than the children, and there is no difference between three- and four-year-olds. Note that children never chose the logically incorrect answer (the Distractor), so we treat chance as 50% rather than 33%.

Previous work has suggested that scalar implicature is difficult for children until quite late in development (Noveck, 2001). Nevertheless three- and four-year-olds performed reliably better than chance in our ad-hoc implicature task which has a similar logical structure to conventional scalar implicatures (see Figure 1). This result suggests that children's difficulties in scalar implicature tasks may not be caused by

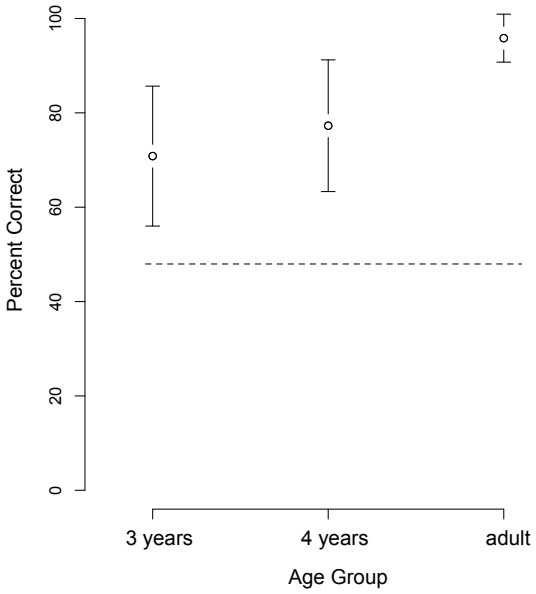

Figure 2: Mean percent correct performance on inference trials for all three age groups in Experiment 1. Error bars show 95% confidence intervals created by a subject-wise nonparametric bootstrap. Dashed line represents chance (50%) even though there were three items, because no child ever chose the logically false distractor. All groups are significantly above chance.

inability to draw pragmatic inferences, but instead by issues with the particular lexical items used in such tasks. This is consistent with findings that children strengthen their understanding of quantifiers when scalar alternatives are made explicitly available (Barner & Bachrach, 2010; Barner, Bale, & Brooks, 2010).

Though children succeeded at our task, adults still outperformed children, possibly reflecting growing pragmatic competence over time. However adults' performance may also reflect explicit strategies: seven (29%) of the responses to our debriefing question "What did you think this study was about?" mentioned scales or the logical/pragmatic dichotomy. Two typical responses were "[the task is to] logically separate items that have more information than given" and "the experiment is to see how literal *[sic]* one takes the instructions."

Experiment 2

What leads to a robust implicature in Experiment 1, even among young children who would fail to draw a conventional implicature? In Experiment 2 (and later Experiment 3) we ask whether the scales which lead to implicature in Experiment 1 are given by the immediate context of objects, or involve additional linguistic or world-knowledge. In Experiment 2 we set up a context which is logically equivalent to the context of Experiment 1, but in which the absence of a feature is replaced with an alternate feature (Figure 1, bottom row).

Table 1: Coefficient estimates from a mixed logistic regression model predicting performance by age group. In the model, 3–4 year-olds are coded as the intercept, thus the coefficients for 4–5 year-olds and adults can be interpreted as tests of whether there is a significant contrast between groups.

	Coef.	Std. Error	z	$p(\|z\|)$
Intercept (3–4 yrs)	0.87	0.35	2.50	0.01
4–5 years	0.43	0.52	0.82	0.41
Adults	2.41	0.65	3.73	$<$0.001

If the inferential effect found in Experiment 1 is derived entirely by the logical alternatives, then we should find the same effect in Experiment 2 as we did in Experiment 1. This would be the prediction from a pure *counterfactual* theory of pragmatic inference: if a speaker wanted to talk about the face with the top hat and glasses, they would have said "top hat"—that they said "glasses" instead indicates that they must have been talking about the other face. We do not expect to see performance above chance, however, if the effect is at least partially driven by linguistic knowledge (e.g. if one doesn't consider alternatives which are more linguistically complex, and therefore "does not have X" isn't an alternative to "has X").

Methods

Participants We posted 28 HITS on Amazon's Mechanical Turk and included 24 responses from participants that gave correct answers on the two filler trials. Participants were paid $0.20 each for completing a HIT.

Stimuli For each of the four trials (faces, houses, pasta, and beds), four separate stimuli were constructed parallel to those in Experiment 1, with the addition of two distinct positive features to replace the absence of features; see Figure 1, bottom row. The stimulus item that previously had no features thus had two novel features, while the stimulus item that previously had one feature now had the old feature as well as one new one. The target stimulus and the position of the answer choices were counterbalanced over six orders.

Procedures Question prompts were identical to those in Experiment 1.

Results and Discussion

In this "no scales" condition, adult participants did not draw an implicature, performing at chance. Results are compared with Experiment 1 in Figure 3. We again used mixed logistic regression to compare adults' performance in Experiment 1 with the performance of the new group in Experiment 2. Coefficient estimates are given in Table 2. The intercept reflects

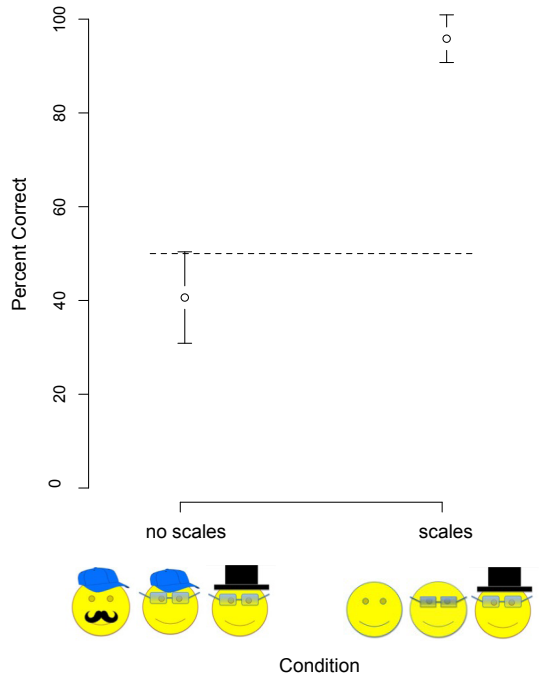

Figure 3: Mean percent correct performance for Experiment 1 (adults) and Experiment 2. Error bars show 95% confidence intervals created by a subject-wise non-parametric bootstrap. Dashed line represents chance (50%).

performance in the "no scales" condition (Experiment 2) and did not differ significantly from zero (indicating chance level responding). In contrast, there was a highly significant coefficient for the "scales" condition (Experiment 1).

In contrast with Experiment 1, participants in Experiment 2 seemed unaware that the task had any pragmatic content. Answers to the debriefing question ranged from "selecting between two objects of significant similarity" to "It was some kind of personality study." Many participants simply stated that they had no idea what the study was about.

The dramatic difference in performance between these two experiments suggests that the logical structure of the immediate context is not enough to lead adults to make a pragmatic inference. Thus, these data cast doubt on a pure counterfactual account of this ad-hoc implicature and suggest that additional knowledge either about the world or about language must be brought to bear.

The linguistic alternatives theory fares better in explaining the disparity between Experiments 1 and 2: for the "scales" condition, the statement "my friend has glasses" is strengthened by negating the alternative "my friend has a top hat", however the negation of the alternative "my friend does not have a top hat" is not included because this is linguistically more complex. For the "no scales" condition *both* "my friend has a top hat" and "my friend has a baseball cap" are negated,

Table 2: Coefficient estimates from a mixed logistic regression predicting adult judgment performance by adults in Experiments 1 and 2. Experiment 2 ("no scales") is coded as the intercept, while a coefficient is fit for Experiment 1 ("scales").

	Coef.	Std. Error	z	$p(\|z\|)$
Intercept (E2)	-0.38	0.23	-1.69	0.09
Scales (E1)	3.53	0.55	6.37	<0.001

leading to chance performance. In Experiment 3 we further test this theory by manipulating participants' world knowledge through changes in the distribution of features like top hat and glasses.

Experiment 3

If the linguistic alternatives theory is a complete explanation of the difference between Experiments 1 and 2, then it should be impossible to make participants succeed in the "no scales" condition without changing the available descriptions. In contrast, if a scale derives from *informativeness* of the possible descriptions then it should be possible to improve performance by providing additional world-knowledge but not altering possible descriptions.

The simple model described in Frank, Goodman, Lai, and Tenenbaum (2009) defined informativeness via information theory: that an informative expression literally conveys more bits of information about which object is being talked about within a context. Although we do not expect that this simple model will capture all the details of Experiments 2 and 3, it does suggest a manipulation to test the pure linguistic alternatives theory: The rarer a feature is, the more informative it is to note that an object has this feature. This principle explains why we would never pick out a person by saying "my friend has legs" (because everyone has legs, so they don't bear mention) but might say "my friend has a mohawk" (because mohawks are rare and hence informative).

This informativeness account can also explain the success of participants in reasoning about the ad-hoc scale in Experiment 1: in an influential analysis of the Wason selection task, Oaksford and Chater (1996) argued that people assume features are rare, and the presence of a feature is more informative than its absence (which explained pervasive reasoning "errors"). The strength of the "scales" inference in Experiment 1 may be due to this rarity assumption setting up a natural informativeness scale.

Experiment 3 tests the hypothesis that manipulating the informativeness of features will allow participants to succeed in the "no scales" condition. We use a pre-exposure to the distribution of features like top hat and glasses to parametrically vary their rarity. We predict that if informativeness sets up orderings (and orderings allow implicature), then as rarity increases, implicatures should increase correspondingly. Put another way: If top hats are rare, speakers who want to talk about people with top hats should mention their hats. Con-

versely, if only the structure of alternative descriptions matter (as in the linguistic alternatives account described above for Experiment 2), then no effect of rarity should be found.

Methods

Participants We posted 344 total HITS to Amazon's Mechanical Turk and received 216 responses (24 per condition) in which participants successfully answered the manipulation check trials (described below) and both filler trials. Participants were compensated $0.20 for completing a HIT.

Stimuli The test stimuli were identical to those in Experiment 2. Familiarization sets of 10 images were constructed, systematically varying the relative frequency of the features. The frequency distribution of objects in the nine conditions is shown at the bottom of Figure 4.

Procedures Experiment 3 was identical to Experiment 2, except that participants viewed a context set that included 10 images immediately before completing each question as in Experiments 1 and 2. They were told that e.g. "In Furble's world there are lots of houses. Here's a picture of the houses." Participants were grouped into distinct between-subjects conditions such that each group saw a different distribution of objects in the context set. Immediately below each context set, participants were asked to select the most frequent object in the set to ensure that they attended to this phase. This question was used as a manipulation check to ensure that participants were paying attention.

Results and Discussion

There was a strong relationship between the frequency of the non-target feature (top hat) and the proportion of participants making implicatures. Results are plotted in Figure 4. We analyzed the data with a mixed linear model, reported in Table 3. The results show that the rarer the omitted portion of an implicature is, the more informative is its omission.

This finding supports the hypothesis outlined above: that the rarer a feature is, the more informative it is and hence the more likely a speaker would be to mention it to pick out a referent. Conversely, when a certain feature is seen as normal, it no longer needs to be included in an informative description. Thus, a failure to mention a feature like "top hat" in a world where top hats are rare strongly implies that the speaker does not want to refer to the face with the top hat. Crucially, this result rules out a simple linguistic alternatives account (as described above): In this experiment, neither the available linguistic descriptions nor their complexity changed as the feature base-rates varied, yet pragmatic inferences varied strikingly.

Participants seemed somewhat aware of the relevant factors underlying Experiment 3, with common debriefing responses taking the form of "you wanted to see if people naturally gravitate to what is most common" and "the object of this study is to determine whether or not the participant understands the

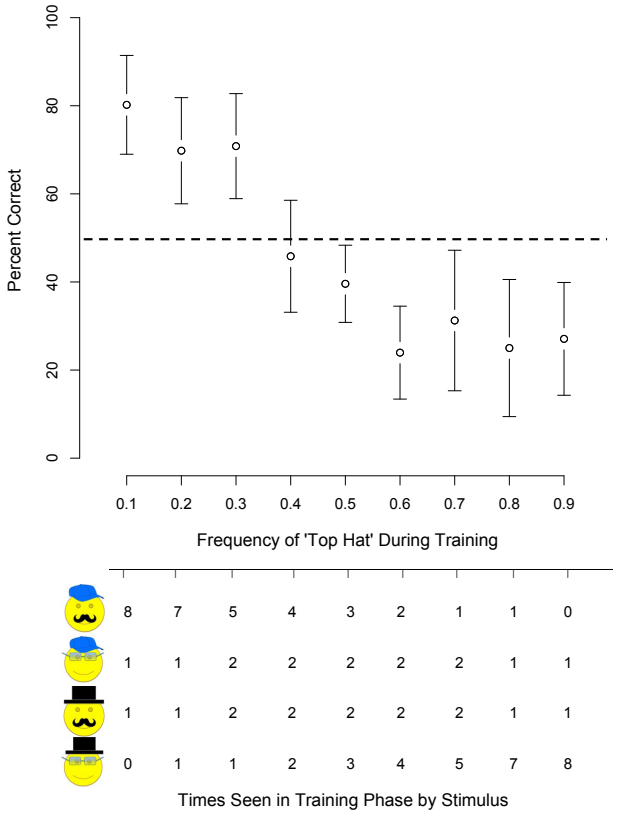

Figure 4: Mean percent correct performance on inference trials for Experiment 1; each point represents a separate condition. The horizontal axis shows the relative frequency of "top hat" (the non-named property) in the context trials; the diagram below shows the number of each object in the familiarization phase. Error bars show 95% confidence intervals by bootstrap, dashed line represents chance.

significance of statistics, and whether or not their understanding of statistics affects their answers to the questions." Nevertheless, it did not seem to be the case that participants were applying simple heuristics related to the presence or absence of features or their conjunctions. For example, the condition with no face with both a top hat and glasses together was no different than the one that included one example of this pair (at .1 and .2 on the horizontal axis in Figure 4, respectively), so the absence of the conjunction did not seem to be critical in participants' responses.

General Discussion

In three experiments we have investigate the nature of scalar implicatures using what we have called ad-hoc scales—scales constructed from contextual, rather than conventional linguistic factors. In contrast to standard scalar implicatures (which are difficult for children below the age of 5), even three-year-olds were able to use ad-hoc scales to disambiguate the refer-

Table 3: Mixed linear model fitting observations in Experiment 3 to frequency of properties during training.

	Coef.	Std. Error	z	$p(\|z\|)$
Intercept	-2.46	0.29	-8.73	$<$0.001
"Top hat" freq.	-4.42	0.49	-9.07	$<$0.001

ent of a logically ambiguous expression (Experiment 1). This results suggests that the inferential mechanisms underlying implicature are present in young children, but, congruent with previous work, children have difficulty construing quantifiers as alternatives in a scale (Barner et al., 2010).

Experiments 2 and 3 contrasted two different theories of the nature of implicature: *counterfactual* and *linguistic alternatives* theories. The Gricean, counterfactual theory predicted that ad-hoc scalar implicatures of the sort described in Experiment 1 would be possible across a range of contexts, even when a featural contrast replaced a contrast between a feature and nothing. The linguistic alternatives theory predicted that the negation of a feature ("no top hat") would be more complex than an alternative feature ("baseball cap") and hence the implicature would be possible in the "scales" condition, but not in the equivalent "no scales" condition. Consistent with the linguistic alternatives theory, participants were at chance in the "no scales" condition (Experiment 2).

Experiment 3 then tested the linguistic alternatives theory more stringently. When participants were trained that certain properties were commonplace and others were informative, they drew strong pragmatic inferences, just as in the "scales" condition. Conversely, when exposed to a distribution in which feature frequencies were flipped, participants made the reverse inferences, as if they had established a scale in the opposite direction. Statistical information about properties closed the gap between the "scales" and "no scales" conditions. These data are inconsistent with a simple linguistic alternatives theory: changing the contextual base rates—without changing the complexity of linguistic alternatives—was enough to invert participants' judgments.

Taken together, these findings support a statistical linguistic account of scalar pragmatics in our ad-hoc task. Adults and children succeeded at the "scales" condition by relying, we suspect, on the real-world knowledge that possessing a feature (e.g. a top hat) is less common than *not* possessing that feature. This statistical information about the rarity and informativeness exemplifies what Sperber and Wilson (1986) call "shared knowledge" and Clark (1996) calls "common ground." A theory of implicature must integrate fine-grained statistical information about such shared context to capture the effects reported here.

Pragmatic computations operate over our knowledge about the world, our knowledge of language, and our knowledge of other people. The research reported here takes a first step towards understanding and manipulating the complex dependencies that go into the computation of pragmatic implicatures.

Acknowledgments

Special thanks to David Barner, Irene Heim, Danny Fox, and Allison Kraus for valuable discussion. Thanks to the staff, teachers, and children at Bing Nursery School for help with Experiment 1.

References

Barner, D., & Bachrach, A. (2010). Inference and exact numerical representation in early language development. *Cognitive Psychology*, *60*(1), 40 - 62.

Barner, D., Bale, A., & Brooks, N. (2010). Quantity implicature and access to scalar alternatives in language acquisition.

Braine, D., & Rumain, B. (1981). Development of comprehension of "or": Evidence for a sequence of competencies. *Journal of Experimental Child Psychology*, *31*(1), 46-70.

Chierchia, G., Fox, D., & Spector, B. (2008). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. *Handbook of Semantics. Mouton de Gruyter, New York, NY*.

Clark, H. (1996). Using language. *Computational Linguistics*, *23*(4).

Fox, D. (2007). Free choice and the theory of scalar implicatures. *Presupposition and implicature in compositional semantics*, 71–120.

Frank, M., Goodman, N., Lai, P., & Tenenbaum, J. (2009). *Informative communication in word production and word learning.*

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models* (Vol. 625). Cambridge University Press Cambridge.

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

Huang, Y., & Snedeker, J. (2009). Semantic meaning and pragmatic interpretation in 5-year-olds: Evidence from real-time spoken language comprehension. *Developmental Psychology*, *45*(6), 1723-1729.

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

Noveck, I. A. (2001). When children are more logical than adults: experimental investigations of scalar implicature. *Cognition*, *78*(2), 165 - 188.

Oaksford, M., & Chater, N. (1996). Rational explanation of the selection task.

Papafragou, A., & Musolino, J. (2003). Scalar implicatures: experiments at the semantics-pragmatics interface. *Cognition*, *86*(3), 253 - 282.

Smith, C. L. (1980). Quantifiers and question answering in young children. *Journal of Experimental Child Psychology*, *30*(2), 191 - 205.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Cambridge, Mass.: Harvard University Press.